Appl. Comput. Harmon. Anal. 34 (2013) 30–56

Contents lists available at SciVerse ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

The cosparse analysis model and algorithms ☆

S. Nam [a,*], M.E. Davies [b], M. Elad [c], R. Gribonval [a]

[a] Centre de Recherche INRIA Rennes – Bretagne Atlantique, Campus de Beaulieu, F-35042 Rennes, France
[b] School of Engineering and Electronics, The University of Edinburgh, Edinburgh, EH9 3JL, UK
[c] Department of Computer Science, The Technion, Haifa 32000, Israel

ARTICLE INFO

Article history:
Received 18 June 2011
Revised 11 December 2011
Accepted 17 March 2012
Available online 21 March 2012
Communicated by Richard Baraniuk

Keywords:
Synthesis
Analysis
Sparse representations
Union of subspaces
Pursuit algorithms
Greedy algorithms
Compressed-sensing

ABSTRACT

After a decade of extensive study of the sparse representation synthesis model, we can safely say that this is a mature and stable field, with clear theoretical foundations, and appealing applications. Alongside this approach, there is an *analysis* counterpart model, which, despite its similarity to the synthesis alternative, is markedly different. Surprisingly, the analysis model did not get a similar attention, and its understanding today is shallow and partial.

In this paper we take a closer look at the analysis approach, better define it as a generative model for signals, and contrast it with the synthesis one. This work proposes effective pursuit methods that aim to solve inverse problems regularized with the analysis-model prior, accompanied by a preliminary theoretical study of their performance. We demonstrate the effectiveness of the analysis model in several experiments, and provide a detailed study of the model associated with the 2D finite difference analysis operator, a close cousin of the TV norm.

1. Introduction

Situated at the heart of signal and image processing, data models are fundamental for stabilizing the solution of inverse problems, and enabling various other tasks, such as compression, detection, separation, sampling, and more. What are those models? Essentially, a model poses a set of mathematical properties that the data is believed to satisfy. Choosing these properties (i.e. the model) carefully and wisely may lead to a highly effective treatment of the signals in question and consequently to successful applications.

Throughout the years, a long series of models has been proposed and used, exhibiting an evolution of ideas and improvements. In this context, the past decade has been certainly the era of sparse and redundant representations, a novel synthesis model for describing signals [5,19,36,51]. Here is a brief description of this model.

Assume that we are to model the signal $\mathbf{x} \in \mathbb{R}^d$. The sparse and redundant synthesis model suggests that this signal could be described as $\mathbf{x} = \mathbf{D}\mathbf{z}$, where $\mathbf{D} \in \mathbb{R}^{d\times n}$ is a possibly redundant dictionary ($n \geqslant d$), and $\mathbf{z} \in \mathbb{R}^n$, the signal's representation, is assumed to be sparse. Measuring the cardinality of non-zeros of $\mathbf{z}$ using the 'ℓ_0-norm', such that $\|\mathbf{z}\|_0$ is the count of the non-zeros in $\mathbf{z}$, we expect $\|\mathbf{z}\|_0$ to be much smaller than n. Thus, the model essentially assumes that any signal from the family of interest could be described as a linear combination of few columns from the dictionary $\mathbf{D}$. The name "synthesis" comes from the relation $\mathbf{x} = \mathbf{D}\mathbf{z}$, with the obvious interpretation that the model describes a way to synthesize a signal.

☆ This work was supported in part by the EU through the project SMALL (Sparse Models, Algorithms and Learning for Large-Scale data), FET-Open programme, under grant number 225913.

* Corresponding author.
E-mail addresses: sangnam.nam@inria.fr (S. Nam), Mike.Davies@ed.ac.uk (M.E. Davies), elad@cs.technion.ac.il (M. Elad), remi.gribonval@inria.fr (R. Gribonval).

1063-5203/$ – see front matter
http://dx.doi.org/10.1016/j.acha.2012.03.006

This model has been the focus of many papers, studying its core theoretical properties by exploring practical numerical algorithms for using it in practice (e.g. [9–11,37]), evaluating theoretically these algorithms' performance guarantees (e.g. [2,15,28,54,55]), addressing ways to obtain the dictionary from a bulk of data (e.g. [1,23,34,47]), and beyond all these, attacking a long series of applications in signal and image processing with this model, demonstrating often state-of-the-art results (e.g. [20,22,32,42]). Today, after a decade of an extensive study along the above lines, with nearly 4000 papers[1] written on this model and related issues, we can safely say that this is a mature and stable field, with clear theoretical foundations, and appealing applications.

Interestingly, the synthesis model has a "twin" that takes an *analysis* point of view. This alternative assumes that for a signal of interest, the analyzed vector $\boldsymbol{\Omega}\mathbf{x}$ is expected to be sparse, where $\boldsymbol{\Omega} \in \mathbb{R}^{p\times d}$ is a possibly redundant *analysis operator* ($p \geqslant d$). Thus, we consider a signal as belonging to the analysis model if $\|\boldsymbol{\Omega}\mathbf{x}\|_0$ is small enough. Common examples of analysis operators include: the shift invariant wavelet transform $\boldsymbol{\Omega}_{\mathrm{WT}}$ [36]; the finite difference operator $\boldsymbol{\Omega}_{\mathrm{DIF}}$, which concatenates the horizontal and vertical derivatives of an image and is closely connected to total variation [45]; the curvelet transform [48], and more. Empirically, analysis models have been successfully used for a variety of signal processing tasks [24,30,48–50] such as denoising, deblurring, and most recently compressed sensing, but this has been done with little theoretical justification.

It is well known by now [21] that for a square and invertible dictionary, the synthesis and the analysis models are the same with $\mathbf{D} = \boldsymbol{\Omega}^{-1}$. The models remain similar for more general dictionaries, although then the gap between them is unexplored. Despite the close-proximity between the two—synthesis and analysis—models, the first has been studied extensively while the second has been left aside almost untouched. In this paper we aim to bring justice to the analysis model by addressing the following set of topics:

1. **Cosparsity:** In Section 2 we start our discussion with a closer look at the sparse analysis model in order to better define it as a generative model for signals. We show that, while the synthesis model puts an emphasis on the non-zeros of the representation vector $\mathbf{z}$, the analysis model draws its strength from the zeros in the analysis vector $\boldsymbol{\Omega}\mathbf{x}$.
2. **Union of subspaces:** Section 2 is also devoted to a comparison between the synthesis model and the analysis one. We know that the synthesis model described above is an instance of a wider family of models, built as a finite union of subspaces [33]. By choosing all the sub-groups of columns from $\mathbf{D}$ that could be combined linearly to generate signals, we get an exponentially large family of low-dimensional subspaces that cover the signals of interest. Adopting this perspective, the analysis model can obtain a similar interpretation. How are the two related to each other? Section 2 considers this question and proposes a few answers.
3. **Uniqueness:** We know that the *spark* of the dictionary governs the uniqueness properties of sparse solutions of the underdetermined linear system $\mathbf{Dz} = \mathbf{x}$ [15]. Can we derive a similar relation for the analysis case? As a platform for studying the analysis uniqueness properties, we consider an inverse problem of the form $\mathbf{y} = \mathbf{Mx}$, where $\mathbf{M} \in \mathbb{R}^{m\times d}$ and $m < d$, and $\mathbf{y} \in \mathbb{R}^m$ is a measurement vector. Put roughly (and this will be better defined later on), assuming that $\mathbf{x}$ comes from the sparse analysis model, could we claim that there is only one possible solution $\mathbf{x}$ that can explain the measurement vector $\mathbf{y}$? Section 3 presents this uniqueness study.
4. **Uniqueness: worked examples:** Based on the study of the analysis uniqueness properties, we characterize the required number of measurements for the uniqueness of the signal that satisfies the analysis model in the case of analysis operator $\boldsymbol{\Omega}$ in general position and the 2D one-step finite difference operator $\boldsymbol{\Omega}_{\mathrm{DIF}}$.
5. **Pursuit algorithms:** Armed with a deeper understanding of the analysis model, we may ask how to efficiently find $\mathbf{x}$ for the above-described linear inverse problem. As in the synthesis case, we can consider either relaxation-based methods or greedy ones. In Section 4 we present two numerical approximation algorithms: a greedy algorithm termed "Greedy Analysis Pursuit" (GAP) that resembles the Orthogonal Matching Pursuit (OMP) [37]—adapted to the analysis model—, and the previously considered ℓ_1-minimization approach [7,21,46]. Section 5 accompanies the presentation of GAP with a theoretical study of its performance guarantee, deriving a condition that resembles the ERC obtained for OMP [54]. Similarly, we study the terms of success of the ℓ_1-minimization approach for the analysis model, deriving a condition that is similar to the one obtained for the synthesis sparse model [54].
6. **Tests:** In Section 6 we demonstrate the effectiveness of the analysis model and the pursuit algorithms proposed in several experiments, starting from synthetic ones (involving random analysis operators) and going all the way to a compressed-sensing test for an image based on the analysis model: the Shepp Logan phantom.

We believe that with the above set of contributions, the cosparse analysis model becomes a well-defined and competitive model to the synthesis counterpart, equipped with all the necessary ingredients for its practical use. Furthermore, this work leads to a series of new questions that are parallel to those studied for the synthesis model—developing novel pursuit methods, a theoretical study of pursuit algorithms for handling other inverse problems, training $\boldsymbol{\Omega}$ just as done for $\mathbf{D}$, and more. We discuss these and other topics in Section 7.

[1] This is a crude estimate, obtained using ISI-Web-of-Science. By first searching Topic = (sparse and representation and (dictionary or pursuit or sensing)), 240 papers are obtained. Then we consider all the papers that cite the above-found, and this results with ≈ 3900 papers.

1.1. Related work

Several works exist in the literature that are related to the analysis model. The work by Elad et al. [21] was the first to observe the dichotomy of analysis and synthesis models for signals. Their study, done in the context of the Maximum-A-Posteriori Probability estimation, presented the two alternatives and explored cases of equivalence between the two. They demonstrated a superiority of the analysis-based approach in signal denoising. Further empirical evidence of the effectiveness of the analysis-based approach for signal and image restoration can be found in [43] and [46]. In [46] it was noted that the non-zero coefficients play a different role in the analysis and synthesis forms but the importance of the zero coefficients for the analysis model—which is reminiscent of signal characterizations through the zero-crossings of their undecimated wavelet transform [35]—was not explicitly identified.

More recently, Candès et al. [7] provided a theoretical study on the error when the analysis-based ℓ_1-minimization is used in the context of compressed sensing. Our work is closely related to these contributions in various ways, and we shall return to these papers when diving into the details of our study.

Some part of this work has been presented in a conference paper [38].

2. A closer look at the cosparse analysis model

We start our discussion with the introduction of the sparse analysis model, and the notion of cosparsity that is fundamental for its definition. We also describe how to interpret the analysis model as a generative one (just like the synthesis counterpart). Finally, we consider the interpretation of the sparse analysis and synthesis models as two manifestations of union-of-subspaces models, and show how they are related.

2.1. Introducing cosparsity

As described in the introduction, a conceptually simple model for data would be to assume that each signal we consider can be expressed (i.e., well-approximated) as a combination of a few building atoms. Once we take this view, a simple synthesis model can be thought of: first, there is a collection of the atomic signals $\{\mathbf{d}_j\}_{j=1}^n \in \mathbb{R}^d$ that we concatenate as the columns of a dictionary, denoted by $\mathbf{D} \in \mathbb{R}^{d\times n}$. Here, typically $n \geqslant d$, implying that the dictionary is redundant. Second, the signal $\mathbf{x} \in \mathbb{R}^d$ can be expressed as a linear combination of some atoms of $\mathbf{D}$, thus there exists $\mathbf{z} \in \mathbb{R}^n$ such that $\mathbf{x} = \mathbf{D}\mathbf{z}$. Third and most importantly, $\mathbf{x}$ must lie in a low-dimensional subspace, and in order to ensure this, very few atoms are used in the expression $\mathbf{x} = \mathbf{D}\mathbf{z}$, i.e., the number of non-zeros $\|\mathbf{z}\|_0$ is very small. By the observation that $\|\mathbf{z}\|_0$ is small, we say that $\mathbf{x}$ has a *sparse representation in* $\mathbf{D}$. The number $k = \|\mathbf{z}\|_0$ is the *sparsity* of the coefficient vector $\mathbf{z}$ and, by extension, of the signal $\mathbf{x}$.

Often, the validity of the above described sparse synthesis model is demonstrated by applying a linear transform to a class of signals to be processed and observing that most of the coefficients are close to zero, exhibiting sparsity. In signal and image processing, discrete transforms such as wavelet, Gabor, curvelet, contourlet, shearlet, and others [14,31,36,48], are of interest, and this empirical observation seems to give a good support for the sparse synthesis model. Indeed, when aiming to claim optimality of a given transform, this is exactly the approach taken—show that for a (theoretically-modeled) class of signals of interest, the transform coefficients tend to exhibit a strong decay. However, one cannot help but noticing that this approach of validating the synthesis model seems to actually validate another 'similar' model; we are considering a model where the signals of interest have *sparse analysis representations*. This point is especially pronounced when the transform used is overcomplete or redundant.

Let us now look more carefully at the above mentioned model that seems to be similar to the sparse synthesis one. First, let $\boldsymbol{\Omega} \in \mathbb{R}^{p\times d}$ be a signal transformation or an *analysis operator*. Its rows are the row vectors $\{\omega_j\}_{j=1}^p$ that will be applied to the signals. Applying $\boldsymbol{\Omega}$ to $\mathbf{x}$, we obtain the (analysis) representation $\boldsymbol{\Omega}\mathbf{x}$ of $\mathbf{x}$. To capture various aspects of the information in $\mathbf{x}$, we typically have $p \geqslant d$.

For simplicity, *unless stated otherwise, we shall assume hereafter that all the rows of* $\boldsymbol{\Omega}$ *are* in general position, i.e., *there are no non-trivial linear dependencies among the rows*.[2] Note that this assumption is used for the purpose of contrasting the analysis model to the synthesis model, and that we will study a case when $\boldsymbol{\Omega}$ is not in general position in later sections. As a matter of fact, there is some indication that linear dependencies among the rows of $\boldsymbol{\Omega}$ can be a 'blessing.' (See, e.g., uniqueness results in Sections 3.3 and 3.4.)

Clearly, unless $\mathbf{x} = 0$, no representation $\boldsymbol{\Omega}\mathbf{x}$ can be 'very sparse,' since at least $p - d$ of the coefficients of $\boldsymbol{\Omega}\mathbf{x}$ are necessarily non-zeros. We shall put our emphasis on the number of zeros in the representation, a quantity we will call *cosparsity*.

Definition 1. The *cosparsity* of a signal $\mathbf{x} \in \mathbb{R}^d$ with respect to $\boldsymbol{\Omega} \in \mathbb{R}^{p\times d}$ (or simply the cosparsity of $\mathbf{x}$) is defined to be:

$$\text{Cosparsity:} \quad \ell := p - \|\boldsymbol{\Omega}\mathbf{x}\|_0. \tag{1}$$

[2] Put differently, we assume that the spark of the matrix $\boldsymbol{\Omega}^T$ is full, implying that every set of d rows from $\boldsymbol{\Omega}$ are linearly independent.

The index set of the zero entries of $\boldsymbol{\Omega}\mathbf{x}$ is called the *cosupport* of $\mathbf{x}$. We say that $\mathbf{x}$ has *cosparse representation* or $\mathbf{x}$ is *cosparse* when the cosparsity of $\mathbf{x}$ is large, where by large we mean that ℓ is close to d. We will see that, while $\ell \leqslant d$ for an analysis operator in general position, there are specific examples where ℓ may exceed d.

At first sight the replacement of *sparsity* by *cosparsity* might appear to be mere semantics. However we will see that this is not the case. In the synthesis model it is the columns $\mathbf{d}_j$, $j \in T$ associated with the index set T of non-zero coefficients that define the signal subspace. Removing columns from $\mathbf{D}$ not in T leaves this subspace unchanged. In contrast, it is the rows ω_j associated with the index set Λ such that $\langle \omega_j, \mathbf{x}\rangle = 0$, $j \in \Lambda$ that define the analysis subspace. In this case removing rows from $\boldsymbol{\Omega}$ for which $\langle \omega_j, \mathbf{x}\rangle \neq 0$ leaves the subspace unchanged.

From this perspective, the cosparse model is more related to signal characterizations from the zero-crossings of their undecimated wavelet transform [35] than to sparse wavelet expansions.

2.2. Sparse analysis model as a generative model

In a Bayesian context, one can think of data models as generators for random signals from a pre-specified probability density function. In that context, the signals that satisfy the k-sparse synthesis model can be generated as follows: first, choose k distinct columns of the dictionary $\mathbf{D}$ at random (e.g. assuming a uniform probability). We denote the index set chosen by T, and clearly $|T| = k$. Second, form a coefficient vector $\mathbf{z}$ that is k-sparse, with zeros outside the support T. The k non-zeros in $\mathbf{z}$ can be chosen at random as well (e.g. Gaussian iid entries). Finally, the signal is created by multiplying $\mathbf{D}$ to the resulting sparse coefficient vector $\mathbf{z}$.

Could we adopt a similar view for the cosparse analysis model? The answer is positive. Similar to the above, one can produce an ℓ-cosparse signal in the following way: first, choose ℓ rows of the analysis operator $\boldsymbol{\Omega}$ at random, and those are denoted by an index set Λ (thus, $|\Lambda| = \ell$). Second, form an arbitrary signal $\mathbf{v}$ in $\mathbb{R}^d$—e.g., a random vector with Gaussian iid entries. Then, project $\mathbf{v}$ onto the orthogonal complement of the subspace generated by the rows of $\boldsymbol{\Omega}$ that are indexed by Λ, this way getting the cosparse signal $\mathbf{x}$. Explicitly, $\mathbf{x} = (\mathbf{Id} - \boldsymbol{\Omega}_\Lambda^T(\boldsymbol{\Omega}_\Lambda\boldsymbol{\Omega}_\Lambda^T)^{-1}\boldsymbol{\Omega}_\Lambda)\mathbf{v}$. Alternatively, one could first find a basis for the orthogonal complement and then generate a random coefficient vector for the basis.

This way, both models can be considered as generators of signals that have a special structure, and clearly, the two signal generators are different. It is now time to ask how those two families of signals inter-relate. In order to answer this question, we take the union-of-subspaces point of view.

2.3. Union-of-subspaces models

It is well known that the sparse synthesis model is a special instance of a wider family of models called *union of subspaces* [4,33]. Given a dictionary $\mathbf{D}$, a vector $\mathbf{z}$ that is exactly k-sparse with support T leads to a signal $\mathbf{x} = \mathbf{D}\mathbf{z} = \mathbf{D}_T\mathbf{z}_T$, a linear combination of k columns from $\mathbf{D}$. The notation $\mathbf{D}_T$ denotes the sub-matrix of $\mathbf{D}$ containing only the columns indexed by T. Denoting the subspace spanned by these columns by $\mathcal{V}_T := \operatorname{span}(\mathbf{d}_j,\ j \in T)$, the sparse synthesis signals belong to the union of all $\binom{n}{k}$ possible subspaces of dimension k,

$$\text{Sparse Synthesis Model:}\quad \mathbf{x} \in \bigcup_{T:\ |T|=k} \mathcal{V}_T. \tag{2}$$

Similarly, the analysis model is associated to a union-of-subspaces model as well. Given an analysis operator $\boldsymbol{\Omega}$, a signal that is exactly ℓ-cosparse with respect to the rows Λ from $\boldsymbol{\Omega}$ is simply in the orthogonal complement to these ℓ rows. Thus, we have[3] $\boldsymbol{\Omega}_\Lambda\mathbf{x} = 0$, which implies that $\mathbf{x} \in \mathcal{W}_\Lambda$, where $\mathcal{W}_\Lambda := \operatorname{span}(\omega_j,\ j \in \Lambda)^\perp = \{\mathbf{x}, \langle \omega_j, \mathbf{x}\rangle = 0,\ \forall j \in \Lambda\}$. Put differently, we may write $\mathcal{W}_\Lambda = \operatorname{Range}(\boldsymbol{\Omega}_\Lambda^T)^\perp = \operatorname{Null}(\boldsymbol{\Omega}_\Lambda)$. Hence, cosparse analysis signals $\mathbf{x}$ belong to the union of all the $\binom{p}{\ell}$ possible such subspaces of dimension $d - \ell$,

$$\text{Cosparse Analysis Model:}\quad \mathbf{x} \in \bigcup_{\Lambda:\ |\Lambda|=\ell} \mathcal{W}_\Lambda. \tag{3}$$

The following table summarizes these two unions of subspaces, where we recall that we assume $\boldsymbol{\Omega}$ and $\mathbf{D}$ in general position.

Model	Subspaces	No. of subspaces	Subsp. dimension
Synthesis	$\mathcal{V}_T := \operatorname{span}(\mathbf{d}_j,\ j \in T)$	$\binom{n}{k}$	k
Analysis	$\mathcal{W}_\Lambda := \operatorname{span}(\omega_j,\ j \in \Lambda)^\perp$	$\binom{p}{\ell}$	$d - \ell$

[3] Note that the notation $\boldsymbol{\Omega}_\Lambda$ refers to restricting *rows* from $\boldsymbol{\Omega}$ indexed by Λ, whereas in the synthesis case we have taken the *columns*. We shall use this convention throughout this paper, where from the context it should be clear whether rows or columns are extracted.

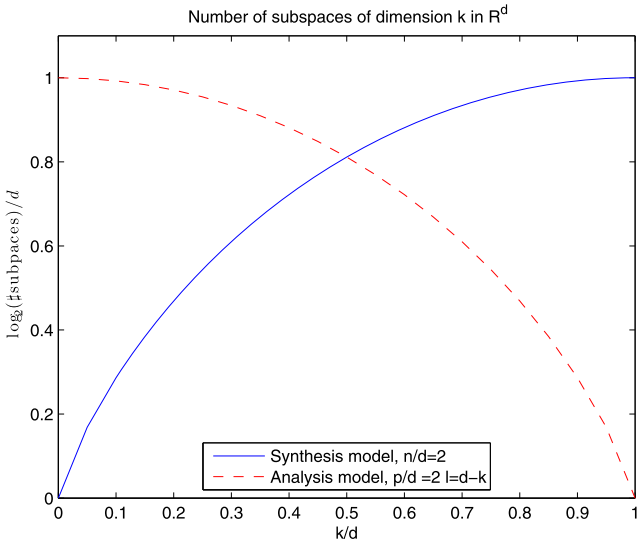

Fig. 1. Number of subspaces of a given dimension, for $n = p = 2d$. The solid (blue) curve shows the log number of subspaces for the synthesis model as the dimension of subspaces vary, while the dashed (red) curve shows that for the analysis model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

What is the relation between these two union of subspaces, as described in Eqs. (2) and (3)? In general, the answer is that the two are different. An interesting way to compare the two models is to consider an ℓ-cosparse analysis model and a corresponding $(d-\ell)$-sparse synthesis model, so that the two have the same dimension in their subspaces.

Following this guideline, we consider first a special case where $\ell = d-1$. In such a case, the dimension of the analysis subspaces is $d-\ell = 1$, and there are $\binom{p}{\ell}$ of those. An equivalent synthesis union of subspaces can be created, where $k = 1$. We should construct a dictionary $\mathbf{D}$ with $n = \binom{p}{\ell}$ atoms $\mathbf{d}_j$, where each atom is the orthogonal complement to one of the sets of ℓ rows from $\boldsymbol{\Omega}$. While the two models become equivalent in this case, clearly $n \gg p$ in general, implying that the sparse synthesis model becomes intractable since $\mathbf{D}$ becomes too large.

By further assuming that $p = d$, we get that there are exactly $\binom{p}{\ell} = \binom{d}{d-1} = d$ subspaces in the analysis union, and in this case $n = p = d$ as well. Furthermore, it is not hard to see that in this case the synthesis atoms are obtained directly by a simple inversion, $\mathbf{D} = \boldsymbol{\Omega}^{-1}$.

Adopting a similar approach, considering the general case where ℓ is a general value (and not necessarily $d-1$), one could always construct a synthesis model that is equivalent to the analysis one. We can compose the synthesis dictionary by simply concatenating all the bases for the orthogonal complements to the subspaces $\mathcal{W}_\Lambda$. The obtained dictionary will have at most $(d-\ell)\binom{p}{\ell}$ atoms. However, not all supports of size k are allowed in the obtained synthesis model, since otherwise the new sparse synthesis model will strictly contain the cosparse analysis one. As such, the cosparse analysis model may be viewed as a sparse synthesis model with some structure.

Further on the comparison between the two models, it would be of benefit to consider again the case $d-\ell = k$ (i.e., having the same dimensionality), assume that $p = n$ (i.e., having the same overcompleteness, for example with $\boldsymbol{\Omega} = \mathbf{D}^T$), and compare the number of subspaces amalgamated in each model. For the sake of simplicity we consider a mild overcompleteness of $p = n = 2d$. Denoting $H(t) := -t\log_2 t - (1-t)\log_2(1-t)$, $0 < t < 1$, the number of subspaces of low-dimension $k \ll d = n/2$ in each data model, from Stirling's approximation, roughly satisfies for large d:

$$\text{Synthesis:}\quad \log_2\binom{n}{k} \approx n \cdot H\left(\frac{k}{n}\right) \approx k \cdot \log_2\frac{n}{k},$$

$$\text{Analysis:}\quad \log_2\binom{p}{\ell} \approx n \cdot H\left(\frac{d-k}{n}\right) \approx n \cdot H(0.5) = n.$$

More generally, unless $d/n \approx 1$, *there are far fewer low-dimensional synthesis subspaces than there are analysis subspaces of the same dimension*. This is illustrated in Fig. 1 when $n = p = 2d$. This indicates a strong difference in the structure of the two models: the synthesis model includes very few low-dimensional subspaces, and an increasingly large number of subspaces of higher dimension, whereas the analysis model contains a combinatorial number of low-dimensional subspaces, with fewer high-dimensional subspaces.

2.3.1. Comment

One must keep in mind that the huge number of low-dimensional subspaces, though rich in terms of its descriptive power, makes it very difficult to recover algorithmically signals that belong to the union of those low-dimensional subspaces

or to efficiently code/sample those signals (see the experimental results in Section 6.1). This stems from the fact that, in general, it is not possible to get cosparsity $d \leqslant \ell < p$: any vector $\mathbf{x}$ that is orthogonal to d linearly independent rows of $\boldsymbol{\Omega}$ must be the zero vector, leading to an uninformative model. One may, however, get cosparsities in the range $d \leqslant \ell < p$ when the analysis operator $\boldsymbol{\Omega}$ displays certain linear dependencies. Therefore it appears to be desirable, in the cosparse analysis model, to have analysis operators that exhibit high linear dependencies among their rows. We will see in Section 3.4 that a leading example of such operators is the finite difference analysis operator.

Another interesting point of view towards the difference between the two models is the following: While a synthesis signal is characterized by the support of the non-zeros in its representation in order to define the subspace it belong to, a signal from the analysis model is characterized by the *locations of the zeros* in its representation $\boldsymbol{\Omega}\mathbf{x}$. The fact that this representation may contain many non-zeroes (and especially so when $p \gg d$) should be of no consequence to the efficiency of the analysis model.

2.4. Comparison with the traditional sparse analysis model

Previous work using analysis representations, both theoretical and algorithmic, has focused on gauging performance in terms of the more traditional sparsity perspective. For example, in the context of compressed sensing, recent theoretical work [7] has provided performance guarantees for minimum ℓ_1-norm analysis representations in this light.

The analysis operator is generally viewed as the dual frame for a redundant synthesis dictionary so that $\boldsymbol{\Omega} = \mathbf{D}^\dagger$. This means that the analysis coefficients $\boldsymbol{\Omega}\mathbf{x}$ provide a consistent *synthesis representation* for $\mathbf{x}$ in terms of the dictionary $\mathbf{D}$, implying that the representation $\boldsymbol{\Omega}\mathbf{x}$ is a feasible solution to the linear system of equations $\mathbf{D}\mathbf{z} = \mathbf{x}$.

Furthermore, if $\|\boldsymbol{\Omega}\mathbf{x}\|_0 = p - \ell$, then $\boldsymbol{\Omega}\mathbf{x}$ must be an element of the k-sparse synthesis model, $\bigcup_{T:|T|=k} \mathcal{V}_T$, with $k = p - \ell$. Hence:

$$\{0\} \subseteq \bigcup_{\Lambda:\ |\Lambda|=p-k} \mathcal{W}_\Lambda \subseteq \bigcup_{T:\ |T|=k} \mathcal{V}_T \subseteq \mathbb{R}^d. \tag{4}$$

Of course, $\boldsymbol{\Omega}\mathbf{x}$ is not guaranteed to be the sparsest representation of $\mathbf{x}$ in terms of $\mathbf{D}$. Hence the two subspace models are not equivalent.

Note that while in Section 2.3 the sparsity k was matched to $d - \ell$, here it is matched to $p - \ell$. The former was used to get the same dimensions in the resulting subspaces, while the match discussed here considers the vector $\boldsymbol{\Omega}\mathbf{x}$ as a candidate k-sparse representation.

Such a perspective treats the analysis operator as a *poor man's* sparse synthesis representation. That is, for certain signals $\mathbf{x}$, the representation $\boldsymbol{\Omega}\mathbf{x}$ may be reasonably sparse but is unlikely to be as sparse as, for example, the minimum ℓ_1-norm synthesis representation.[4]

In the context of linear inverse problems, it is tempting to try to exploit the nesting property (4) in order to derive identifiability guarantees in terms of the sparsity of the analysis coefficients $\boldsymbol{\Omega}\mathbf{x}$. For example, in [7], the compressed sensing recovery guarantees exploit the nesting property (4) by assuming a sufficient number of observations to achieve a stable embedding (restricted isometry property) for the k-sparse synthesis union of subspaces, which in turn implies a stable embedding of the $(p-k)$-cosparse analysis union of subspaces.

While such an approach is of course valid, it misses a crucial difference between the analysis and synthesis representations: they do not correspond to equivalent signal models. Treating the two models as equivalent hides the fact that they may be composed of subspaces with markedly different dimensions. The difference between these models is highlighted in the following examples.

2.4.1. Example: generic analysis operators, $p = 2d$

Assuming the rows of $\boldsymbol{\Omega}$ are in general position, then when $p \geqslant 2d$ the nesting property (4) is trivial but rather useless! Indeed, if $k < d$, then the only analysis signal for which $\|\boldsymbol{\Omega}\mathbf{x}\|_0 = k = p - \ell$ is $\mathbf{x} = 0$. Alternatively, if $k \geqslant d$, the synthesis model is trivially the full space: $\bigcup_{T:\ |T|=k} \mathcal{V}_T = \mathbb{R}^d$.

2.4.2. Example: shift invariant wavelet transform

The shift invariant wavelet transform is a popular analysis transform in signal processing. It is particularly good for processing piecewise smooth signals. Its inverse transform has a synthesis interpretation as the redundant wavelet dictionary consisting of wavelet atoms with all possible shifts.

The shift invariant wavelet transform [36] provides a nice example of an analysis operator that has significant dependencies due to the finite support of the individual wavelets. Such non-trivial dependencies within the rows of $\boldsymbol{\Omega}_{\mathrm{WT}}$ mean that the dimensions of the (analysis or synthesis) signal subspaces are not easily characterized by either the sparsity k or the cosparsity ℓ. However the behavior of the model is still driven by the zero coefficients not the non-zero ones, i.e., by

[4] When measuring sparsity with an ℓ_p norm, $0 < p \leqslant 1$, rather than with $p = 0$, it has been shown [29] that for so-called *localized frames* the analysis coefficients $\boldsymbol{\Omega}\mathbf{x}$ obtained with $\boldsymbol{\Omega} = \mathbf{D}^\dagger$ the canonical dual frame of $\mathbf{D}$ are near optimally sparse: $\|\boldsymbol{\Omega}\mathbf{x}\|_p \leqslant C_p \min_{\mathbf{z}|\mathbf{D}\mathbf{z}=\mathbf{x}} \|\mathbf{z}\|_p$, where the constant C_p does not depend on $\mathbf{x}$.

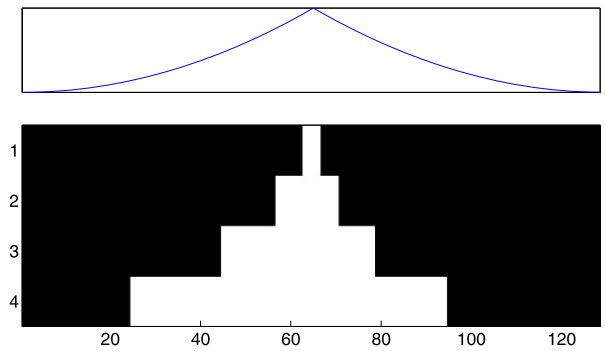

Fig. 2. Top: a piecewise quadratic signal. Bottom: the support set (white region) for the wavelet coefficients of the signal using a $J = 4$ level shift invariant Daubechies wavelet transform with $s = 3$ vanishing moments. Scaling coefficients are not shown. The support set contains 122 coefficients out of a possible 512, yet the analysis subspace has a dimension of only 3.

the zero-crossings of the wavelet transform [35]. By considering a particular support set of an analysis representation $\boldsymbol{\Omega}_{\mathrm{WT}}\mathbf{x}$ with the shift invariant wavelet transform we can illustrate the dramatic difference between the analysis and synthesis interpretations of the coefficients.

Fig. 2 shows the support set of the non-zero analysis coefficients (white region), associated with the cone of influence around a discontinuity in a piecewise polynomial signal of length 128-samples [18], using a shift invariant Daubechies wavelet transform with $s = 3$ vanishing moments [36]. For such a signal, the cone of influence at level J in a shift invariant wavelet transform contains $L_j - 1$ non-zero coefficients where L_j is the length of the wavelet filter at level j. Note though, the non-zero coefficients are not linearly independent and can be elegantly described through the notion of wavelet footprints [18].

2.4.2.1. Synthesis perspective Interpreting the support set within the synthesis model implies that the signal is not particularly sparse and needs a significant number of wavelet atoms to describe it: in Fig. 2 the size of the support set, excluding coefficients of scaling functions, is 122. Could the support set be significantly reduced by using a better support selection strategy such as ℓ_1 minimization? In practice, using ℓ_1 minimization, a support set of 30 can be obtained, again ignoring scaling coefficients.

2.4.2.2. Analysis perspective The analysis interpretation of the shift invariant wavelet representation relies on the examination of the size of the analysis subspace associated with the cosupport set. From the theory of wavelet footprints, the dimension of this subspace is equal to the number of vanishing moments of the wavelet filter, which in this example is only … 3, providing a much lower-dimensional signal model.

We therefore see that the analysis model has a much lower number of degrees of freedom for this support set, leading to a significantly more parsimonious model.

2.5. Hybrid analysis/synthesis models?

In this section we have demonstrated that while both the cosparse analysis model and the sparse synthesis model can be described by a union of subspaces these models are typically very different. We do not argue that one is inevitably better than the other. The value of the model will very much depend on the problem instance. Indeed the intrinsic difference between the models also suggests that it might be fruitful to explore building other union-of-subspace models from hybrid compositions of analysis and synthesis operators. For example, one could imagine a signal model where $\mathbf{x} = \mathbf{D}\mathbf{z}$ through a redundant synthesis dictionary but instead of imposing sparsity on $\mathbf{z}$ we restrict $\mathbf{z}$ through an additional analysis operator: $\|\boldsymbol{\Omega}\mathbf{z}\|_0 \leqslant k$. In such a case there will still be an underlying union-of-subspace model but with the subspaces defined by a combination of atoms and analysis operator constraints. A special case of this is the split analysis model suggested in [7].

3. Uniqueness properties

In the synthesis model, if a dictionary $\mathbf{D}$ is redundant, then a given signal $\mathbf{x}$ can admit many synthesis representations $\tilde{\mathbf{z}}$, i.e., $\tilde{\mathbf{z}}$ with $\mathbf{D}\tilde{\mathbf{z}} = \mathbf{x}$. This makes the following type of problem interesting in the context of the sparse signal recovery: when a signal has a sparse representation $\mathbf{z}$, can there be another representation that is equally sparse or sparser? This problem is well-understood in terms of the so-called *spark* of $\mathbf{D}$ [15], the smallest number of columns from $\mathbf{D}$ that are linearly dependent.

Unlike in the synthesis model, if the signal is known, then its analysis representation $\boldsymbol{\Omega}\mathbf{x}$ with respect to an analysis operator $\boldsymbol{\Omega}$ is completely determined. Hence, there is no inherent question of uniqueness for the cosparse analysis model. The uniqueness question we want to consider in this paper is in the context of the noiseless linear inverse problem,

$$\mathbf{y} = \mathbf{M}\mathbf{x}, \tag{5}$$

where $\mathbf{M} \in \mathbb{R}^{m\times d}$, and $m < d$, implying that the measurement vector $\mathbf{y} \in \mathbb{R}^m$ is not sufficient to fully characterize the original signal $\mathbf{x} \in \mathbb{R}^d$. For this problem we ask: when can we assert that a solution $\mathbf{x}$ with cosparsity ℓ is the only solution with that cosparsity or more? The problem (5) (especially, with additive noise) arises ubiquitously in many applications, and we shall focus on this problem throughout this paper as a platform for introducing the cosparse analysis model, its properties and behavior. Not to complicate matters unnecessarily, we assume that all the rows of $\mathbf{M}$ are linearly independent, and we omit noise, leaving robustness analysis to further work.

For completeness of our discussion, let us return for a moment to the synthesis model and consider the uniqueness property for the inverse problem posed in Eq. (5). Assuming that the signal's sparse representation satisfies $\mathbf{x} = \mathbf{D}\mathbf{z}$, we have that $\mathbf{y} = \mathbf{M}\mathbf{x} = \mathbf{M}\mathbf{D}\mathbf{z}$. Had we known the support T of $\mathbf{z}$, this linear system would have reduced to $\mathbf{y} = \mathbf{M}\mathbf{D}_T\mathbf{z}_T$, a system of m equations with k unknowns. Thus, recovery of $\mathbf{x}$ from $\mathbf{y}$ is possible only if $k \leqslant m$.

When the support of $\mathbf{z}$ is unknown, it is the *spark* of the compound matrix $\mathbf{M}\mathbf{D}$ that governs whether the cardinality of $\mathbf{z}_T$ is sufficient to ensure uniqueness—if $k = \|\mathbf{z}\|_0$ is smaller than half the *spark* of $\mathbf{M}\mathbf{D}$, then necessarily $\mathbf{z}$ is the signal's sparsest representation. At best, $\operatorname{Spark}(\mathbf{M}\mathbf{D}) = m+1$, and then we require that the number of measurements is at least twice the cardinality k. Put formally, we require

$$k = \|\mathbf{z}\|_0 < \frac{1}{2}\operatorname{Spark}(\mathbf{M}\mathbf{D}) \leqslant \frac{m+1}{2}. \tag{6}$$

It will be interesting to contrast this requirement with the one we will derive hereafter for the analysis model.

3.1. Uniqueness when the cosupport is known

Before we tackle the uniqueness problem for the analysis model, let us consider an easier question: given the observations $\mathbf{y}$ obtained via a measurement matrix $\mathbf{M}$, and assuming that the cosupport Λ of the signal $\mathbf{x}$ is known, what are the sufficient conditions for the recovery of $\mathbf{x}$? The answer to this question is straightforward since $\mathbf{x}$ satisfies the linear equation

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix} = \begin{bmatrix}\mathbf{M}\\ \boldsymbol{\Omega}_\Lambda\end{bmatrix}\mathbf{x} = \mathbf{A}\mathbf{x}. \tag{7}$$

To be able to uniquely identify $\mathbf{x}$ from Eq. (7), the matrix $\mathbf{A}$ must have a zero null space. This is equivalent to the requirement

$$\operatorname{Null}(\boldsymbol{\Omega}_\Lambda) \cap \operatorname{Null}(\mathbf{M}) = \mathcal{W}_\Lambda \cap \operatorname{Null}(\mathbf{M}) = \{0\}. \tag{8}$$

Let us now assume that $\mathbf{M}$ and $\boldsymbol{\Omega}$ are mutually independent, in the sense that there are no non-trivial linear dependencies among the rows of $\mathbf{M}$ and $\boldsymbol{\Omega}$; this is a reasonable assumption because first, one should not be measuring something that may be already available from $\boldsymbol{\Omega}$, and second, for a fixed $\boldsymbol{\Omega}$, mutual independency holds true for almost all $\mathbf{M}$ (in the Lebesgue measure). Then, (8) would be satisfied as soon as $\dim(\mathcal{W}_\Lambda) + \dim(\operatorname{Null}(\mathbf{M})) \leqslant d$, or $\dim(\mathcal{W}_\Lambda) \leqslant m$, since $\dim(\operatorname{Null}(\mathbf{M})) = d - m$. This motivates us to define

$$\kappa_{\boldsymbol{\Omega}}(\ell) := \max_{|\Lambda|\geqslant\ell} \dim(\mathcal{W}_\Lambda). \tag{9}$$

The quantity $\kappa_{\boldsymbol{\Omega}}(\ell)$ plays an important role in determining the necessary and sufficient cosparsity level for the identification of cosparse signals. Indeed, under the assumption of the mutual independence of $\boldsymbol{\Omega}$ and $\mathbf{M}$, a necessary and sufficient condition for the uniqueness of *every* cosparse signal given the knowledge of its cosupport Λ of size ℓ is

$$\kappa_{\boldsymbol{\Omega}}(\ell) \leqslant m. \tag{10}$$

3.2. Uniqueness when the cosupport is unknown

The uniqueness question that we answered above refers to the case where the cosupport is known, but of course, in general this is not the case. We shall assume that we may only know the cosparsity level ℓ, which means that our uniqueness question now becomes: what cosparsity level ℓ guarantees that there can be only one signal $\mathbf{x}$ matching a given observation $\mathbf{y}$?

As we have seen, the cosparse analysis model is a special case of a general union-of-subspaces model. Uniqueness guarantees for missing data problems such as (5) with general union-of-subspace models are covered in [4,33]. In particular [33] shows that $\mathbf{M}$ is invertible on the union-of-subspaces $\bigcup_{\gamma\in\Gamma} S_\gamma$ *if and only if* $\mathbf{M}$ is invertible on all subspaces $S_\gamma + S_\theta$ for all $\gamma, \theta \in \Gamma$. In the context of the analysis model this gives the following result whose proof is a direct consequence of the results in [33]:

Proposition 2. *(See [33].) Let $\bigcup_\Lambda \mathcal{W}_\Lambda$, $|\Lambda| = \ell$ be the union of ℓ-cosparse analysis subspaces induced by the analysis operator $\boldsymbol{\Omega}$. Then the following statements are equivalent*:

1. *If the linear system* $\mathbf{y} = \mathbf{M}\mathbf{x}$ *admits an* ℓ*-cosparse solution, then this is the unique* ℓ*-cosparse solution*;
2. $\mathbf{M}$ *is invertible on* $\bigcup_{\Lambda} \mathcal{W}_{\Lambda}$;
3. $(\mathcal{W}_{\Lambda_1} + \mathcal{W}_{\Lambda_2}) \cap \mathrm{Null}(\mathbf{M}) = 0$ *for any* $|\Lambda_1|, |\Lambda_2| \geqslant \ell$.

Proposition 2 answers the question of uniqueness for cosparse signals in the context of linear inverse problems. Unfortunately, the answer we obtained still leaves us in the dark in terms of the necessary cosparsity level or necessary number of measurements. In order to pose a clearer condition, we use Proposition 2 from [33] that poses a sharp condition on the number of measurements to guarantee uniqueness (when $\mathbf{M}$ and $\boldsymbol{\Omega}$ are mutually independent):

$$m \geqslant \tilde{\kappa}_{\boldsymbol{\Omega}}(\ell), \quad \text{where } \tilde{\kappa}_{\boldsymbol{\Omega}}(\ell) := \max\left\{\dim(\mathcal{W}_{\Lambda_1} + \mathcal{W}_{\Lambda_2}) \colon |\Lambda_i| \geqslant \ell,\ i = 1, 2\right\}. \tag{11}$$

Interestingly, a sufficient condition can also be obtained using the quantity $\kappa_{\boldsymbol{\Omega}}$ defined in (9) above, which was observed to play an important role in the uniqueness result when the cosupport is assumed to be known. Namely, we have the following result.

Proposition 3. *Assume that* $\kappa_{\boldsymbol{\Omega}}(\ell) \leqslant \frac{m}{2}$. *Then for almost all* $\mathbf{M}$ (*w.r.t. the Lebesgue measure*), *the linear inverse problem* $\mathbf{y} = \mathbf{M}\mathbf{x}$ *has at most one* ℓ*-cosparse solution.*

Proof. Assuming the mutual independence of $\boldsymbol{\Omega}$ and $\mathbf{M}$, which holds for almost all $\mathbf{M}$, we note that the uniqueness of ℓ cosparse solutions holds if and only if: $\dim(\mathcal{W}_{\Lambda_1} + \mathcal{W}_{\Lambda_2}) \leqslant m$, whenever $|\Lambda_i| \geqslant \ell$, $i = 1, 2$. Assume that $\kappa_{\boldsymbol{\Omega}}(\ell) \leqslant m/2$. By definition of $\kappa_{\boldsymbol{\Omega}}$, if $|\Lambda_i| \geqslant \ell$, $i = 1, 2$, then $\dim(\mathcal{W}_{\Lambda_i}) \leqslant \frac{m}{2}$, hence $\dim(\mathcal{W}_{\Lambda_1} + \mathcal{W}_{\Lambda_2}) \leqslant m$. □

In the synthesis model the degree to which columns are interdependent can be partially characterized by the *spark* of $\mathbf{D}$ [15] defined as the smallest number of columns of $\mathbf{D}$ that are linearly dependent. Here the function $\kappa_{\boldsymbol{\Omega}}$ plays a similar role in quantifying the interdependence between rows in the analysis model.

Remark 4. The condition $\kappa_{\boldsymbol{\Omega}}(\ell) \leqslant \frac{m}{2}$ is in general not necessary while condition (11) is.

There are two classes of analysis operators for which the function $\kappa_{\boldsymbol{\Omega}}$ is well-understood: analysis operators in general position and the finite difference operators. We discuss the uniqueness results for these two classes in the following subsections.

3.3. Analysis operators in general position

It can be easily checked that $\kappa_{\boldsymbol{\Omega}}(\ell) = \max(d - \ell, 0)$. This enables us to quantify the exact level of cosparsity necessary for the uniqueness guarantees:

Corollary 5. *Let* $\boldsymbol{\Omega} \in \mathbb{R}^{p \times d}$ *be an analysis operator* in general position. *Then, for almost all* $m \times d$ *matrices* $\mathbf{M}$, *the following hold*:

- *Based on Eq.* (10), *if* $m \geqslant d - \ell$, *then the equation* $\mathbf{y} = \mathbf{M}\mathbf{x}$ *has at most one solution with known cosupport* Λ (*of cosparsity at least* ℓ).
- *Based on Proposition* 2, *if* $m \geqslant 2(d - \ell)$, *then the equation* $\mathbf{y} = \mathbf{M}\mathbf{x}$ *has at most one solution with cosparsity at least* ℓ.

3.4. The finite difference operator

An interesting class of analysis operators with significant linear dependencies is the family of finite difference operators on graphs, $\boldsymbol{\Omega}_{\mathrm{DIF}}$. These are strongly related to TV norm minimization, popular in image processing applications [45], and has the added benefit that we are able to quantify the function $\kappa_{\boldsymbol{\Omega}}$ and hence the uniqueness properties of the cosparse signal model under $\boldsymbol{\Omega}_{\mathrm{DIF}}$.

We begin by considering $\boldsymbol{\Omega}_{\mathrm{DIF}}$ on an arbitrary graph before restricting our discussion to the 2D lattice associated with image pixels. Consider a non-oriented graph with vertices V and edges $E \subset V^2$. An edge e is a pair $e = (v_1, v_2)$ of connected vertices. For any vector of coefficients defined on the vertices, $\mathbf{x} \in \mathbb{R}^V$, the finite difference analysis operator $\boldsymbol{\Omega}_{\mathrm{DIF}}$ computes the collection of differences $(x(v_1) - x(v_2))$ between end-points, for all edges in the graph. Thus, an edge $e \in E$ may be viewed as a finite difference on $\mathbb{R}^V$.

Can we estimate the function $\kappa_{\boldsymbol{\Omega}_{\mathrm{DIF}}}(\ell)$? The following shows that it is intimately related to topological properties of the graph. For each sub-collection $\Lambda \subset E$ of edges, we can define its vertex-set $V(\Lambda) \subset V$ as the collection of vertices covered by at least one edge in Λ. The support set $V(\Lambda)$ of Λ can be decomposed into $J(\Lambda)$ connected components (a connected component is a set of vertices connected to one another by a walk through vertices in Λ). It is easy to check that a vector $\mathbf{x}$ belongs to the space $\mathcal{W}_{\Lambda} = \mathrm{Null}(\boldsymbol{\Omega}_{\Lambda})$ if and only if its values are constant on each connected component. As a result, the

Fig. 3. Top left: an example of a piecewise constant image: the 256×256 Shepp Logan phantom. Top right: an image with the same cosparsity, $\ell = 128\,014$, but whose cosupport is associated with an empirically maximum subspace dimension. Bottom: zoom on the top of the top right image.

dimension of this subspace is given by

$$\dim(\mathcal{W}_\Lambda) = |V| - \big|V(\Lambda)\big| + J(\Lambda),$$

where the $|V| - |V(\Lambda)|$ vertices out of V are associated to arbitrary values in $\mathbf{x}$ that are distinct from all their neighbors, while all entries from each of the $J(\Lambda)$ connected components have an arbitrary common value. It follows that

$$\kappa_{\boldsymbol{\Omega}}(\ell) = \max_{|\Lambda| \geqslant \ell} \big\{ |V| - \big|V(\Lambda)\big| + J(\Lambda) \big\} = |V| - \min_{|\Lambda| \geqslant \ell} \big\{ \big|V(\Lambda)\big| - J(\Lambda) \big\}. \tag{12}$$

Because of the nesting of the subspaces $\mathcal{W}_\Lambda$, the minimum on the right-hand side is achieved when $|\Lambda| = \ell$.

Uniqueness condition for cosparse images with respect to the 2D $\boldsymbol{\Omega}_{\mathrm{DIF}}$ In the abstract context of general graphs, the characterization (12) may remain obscure, but can we get more concrete estimates by specializing to the 2D regular graph associated to the pixels of an $N \times N$ image? It turns out that one can obtain relatively simple upper and lower bounds for $\kappa_{\boldsymbol{\Omega}_{\mathrm{DIF}}}$ and hence derive an easily interpretable uniqueness condition (see Appendix C for a proof):

Proposition 6. *Let $\boldsymbol{\Omega}_{\mathrm{DIF}}$ be the* 2D *finite difference analysis operator that computes horizontal and vertical discrete derivatives of a $d = N \times N$ image. For any $\ell \geqslant 5$ we have*

$$d - \frac{\ell}{2} - \sqrt{\frac{\ell}{2}} - 1 \leqslant \kappa_{\boldsymbol{\Omega}_{\mathrm{DIF}}}(\ell) \leqslant d - \frac{\ell}{2} - \sqrt{\frac{\ell}{2}} + \frac{1}{2}. \tag{13}$$

As a result, assuming that $\mathbf{M}$ is 'mutually independent' from $\boldsymbol{\Omega}_{\mathrm{DIF}}$, and if

$$m \geqslant 2d - \ell - \sqrt{2\ell} + \frac{1}{2} \geqslant 2\kappa_{\boldsymbol{\Omega}_{\mathrm{DIF}}}(\ell) \tag{14}$$

then the equation $\mathbf{y} = \mathbf{M}\mathbf{x}$ has at most one solution with cosparsity at least ℓ.

The 2D $\boldsymbol{\Omega}_{\mathrm{DIF}}$, piecewise constant images, and the TV norm The 2D finite difference operator is closely related to the TV norm [45]: the discrete TV norm of $\mathbf{x}$ is essentially a mixed $\ell_2 - \ell_1$ norm of $\boldsymbol{\Omega}_{\mathrm{DIF}}\mathbf{x}$. Just like its close cousin TV norm minimization, the minimization of $\|\boldsymbol{\Omega}\mathbf{x}\|_0$ is particularly good at inducing piecewise constant images. We illustrate this through a worked example.

Consider the popular Shepp Logan phantom image shown in the left-hand side of Fig. 3. We denote the cosupport of the image by Λ in line with the discussion in this section: an edge belongs to Λ if a pair of horizontally or vertically neighboring pixels v_1 and v_2 have the same value. This particular image has 14 distinct connected regions of constant intensity. The number of non-zero coefficients in the finite difference representation is determined by the total length (Manhattan distance) of the boundaries between these regions. For the Shepp Logan phantom this length is 2546 pixel widths and thus the cosparsity is $\ell = 130\,560 - 2546 = 128\,014$. Furthermore, as there are no isolated pixels with any other intensity, all pixels belong to a constant intensity region so that $|V(\Lambda)| = |V|$ and the cosupport has an associated subspace dimension of:

$$\dim(\mathcal{W}_\Lambda) = \big(|V| - \big|V(\Lambda)\big|\big) + J(\Lambda) = 14.$$

In order to determine when the Shepp Logan image is the unique solution to $\mathbf{y} = \mathbf{M}\mathbf{x}$ with maximum cosparsity it is necessary to consider the maximum subspace dimension of all possible support sets with the same cosparsity. This is the quantity measured by $\kappa_{\boldsymbol{\Omega}_{\mathrm{DIF}}}(\ell)$.

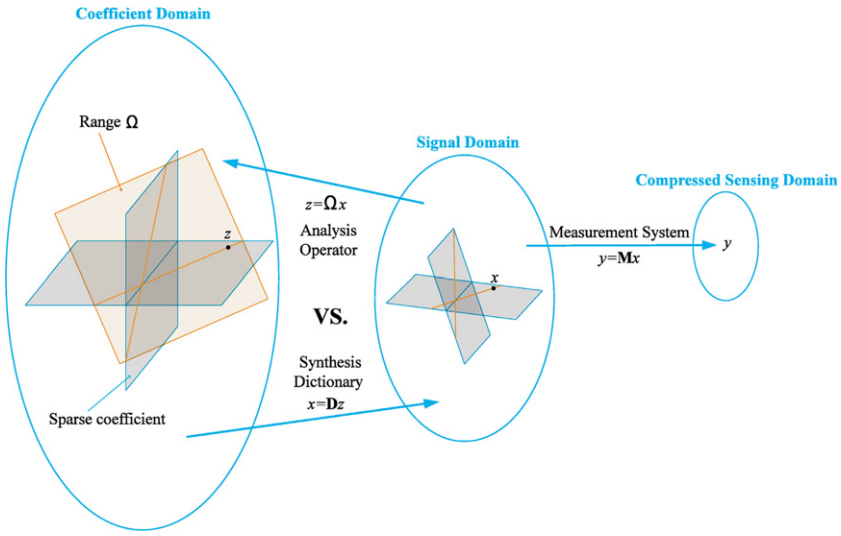

Fig. 4. A schematic overview of analysis cosparse *vs.* synthesis sparse models in relation with compressed sensing.

Following the arguments used in the proof of Proposition 6 we need to find an image for which the $\boldsymbol{\Omega}_{\mathrm{DIF}}$ cosupport, Λ, is a single connected subgraph that is as close to a square as possible. Such an image is shown in the right-hand side of Fig. 3. For this image we have $\dim(\mathcal{W}_\Lambda) = 1276$. Comparing this to the bounds given in (13) of Proposition 6

$$1275 \leqslant \kappa_{\boldsymbol{\Omega}_{\mathrm{DIF}}}(\ell) \leqslant 1276,$$

we see that in this instance the upper bound has been achieved. The uniqueness result from Proposition 6 then tells us that a sufficient number of measurements to uniquely determine the Shepp Logan image is given by $m = 2\kappa_{\boldsymbol{\Omega}_{\mathrm{DIF}}}(128014) = 2552$.

We will revisit this again in Section 6.2 where we investigate the empirical recovery performance of some practical reconstruction algorithms.

3.5. *Overview of cosparse* vs *sparse models for inverse problems*

To conclude this section, Fig. 4 provides a schematic overview of analysis cosparse models *vs.* synthesis sparse models in the context of linear inverse problems such as compressed sensing. In the synthesis model, the signal $\mathbf{x}$ is a projection (through the dictionary $\mathbf{D}$) of a high-dimensional vector $\mathbf{z}$ living in the union of sparse coefficient subspaces; in the analysis model, the signal lives in the pre-image by the analysis operator $\boldsymbol{\Omega}$ of the intersection between the range of $\boldsymbol{\Omega}$ and this union of subspaces. For a given sparsity of $\mathbf{z}$, this is usually a set of much smaller dimensionality.

4. Pursuit algorithms

Having a theoretical foundation for the uniqueness of the problem

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\boldsymbol{\Omega}\mathbf{x}\|_0 \quad \text{subject to} \quad \mathbf{M}\mathbf{x} = \mathbf{y}, \tag{15}$$

we now turn to the question of how to solve it: algorithms. We present two algorithms, both targeting the solution of problem (15). As in the uniqueness discussion, we assume that $\mathbf{M} \in \mathbb{R}^{m\times d}$, where $m < d$. This implies that the equation $\mathbf{M}\mathbf{x} = \mathbf{y}$ has infinitely many possible solutions, and the term $\|\boldsymbol{\Omega}\mathbf{x}\|_0$ introduces the analysis model to regularize the problem.

4.1. *The cosparse signal recovery problem is NP-complete*

Related to (15), we can consider a cosparse signal recovery problem COSPARSE consisting of a quintuplet $(\mathbf{y}, \mathbf{M}, \boldsymbol{\Omega}, \ell, \epsilon)$ in which we seek to find a vector $\mathbf{x}^*$ that satisfies

$$\left\|\mathbf{y} - \mathbf{M}\mathbf{x}^*\right\|_2 \leqslant \epsilon, \qquad \left\|\boldsymbol{\Omega}\mathbf{x}^*\right\|_0 \leqslant p - \ell, \tag{16}$$

where p is the number of rows of $\boldsymbol{\Omega}$ as before. It is easy to see that the decision problem "given $(\mathbf{y}, \mathbf{M}, \boldsymbol{\Omega}, \ell, \epsilon)$, does there exist $\mathbf{x}^*$ satisfying (16)?" is NP [25]: given a candidate solution, one can check in polynomial time whether it satisfies the constraints (16). Moreover, every instance of the classical NP-complete (ϵ, k) SPARSE approximation problem [13,40] can trivially be reduced to an instance of the above decision problem with $\boldsymbol{\Omega} = \mathbf{Id}$, hence COSPARSE is indeed NP-complete.

The above consideration prompts us to look for ways to solve (15) in an 'approximate' way. We discuss two possibilities, a convex relaxation and a greedy approach, with an emphasis on the latter. Of course, there can be many more possibilities to solve (15) or to find approximate solutions of it. We mention a few works where some of such methods can be found in [6,43,46].

4.2. The analysis ℓ_1-minimization

A natural convex relaxation of (15) is to solve:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\boldsymbol{\Omega}\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{M}\mathbf{x} = \mathbf{y}. \tag{17}$$

The analysis ℓ_1-minimization is well known and widely used already in practice (see e.g. [22,51]). The attractiveness of this approach comes from that the convexity of (17) admits computationally tractable algorithms to solve the problem, and that the ℓ_1-norm promotes high cosparsity in the solution $\hat{\mathbf{x}}$. An algorithm that targets the solution of (17) and its convergence analysis can be found in [6]. There are many other papers that have introduced algorithms to solve problems of the form (17) or variants thereof. To give just an example, [44] proposes a general form of forward–backward splitting that can be exploited to deal with such problems.

4.3. The Greedy Analysis Pursuit (GAP) algorithm

The algorithm we present in this section is named Greedy Analysis Pursuit (GAP). It is a variant of well-known greedy pursuit algorithm used for the synthesis model—the Orthogonal Matching Pursuit (OMP) algorithm. Similar to the synthesis case, our goal is to detect the informative support of $\boldsymbol{\Omega}\mathbf{x}$—as discussed in Section 3.1, in the analysis case, this amounts to the locations (cosupport) of the zeros in the vector $\boldsymbol{\Omega}\mathbf{x}$, so as to introduce additional constraints to the underdetermined system $\mathbf{M}\mathbf{x} = \mathbf{y}$. Note that for obtaining a solution, one needs to detect at least $d - m$ of these zeros, and thus if $\ell > d - m$, detection of the complete set of zeros is not mandatory.

An obvious way to find the cosupport of a cosparse signal would proceed as follows: first, obtain a reasonable estimate of the signal from the given information. Using the initial estimate, select a location as belonging to the cosupport. Having this estimated part of the cosupport, we can obtain a new estimate. One can now see that by alternating the two previous steps, we will have estimated enough locations of the cosupport to get the final estimate.

However, the GAP works in an opposite direction and aims to detect the elements *outside* the set Λ, this way carving its way towards the detection of the desired cosupport. Therefore, the cosupport $\hat{\Lambda}$ is initialized to be the whole set $\{1, 2, 3, \ldots, p\}$, and through the iterations it is reduced towards a set of size ℓ (or less, $d - m$).

Let us discuss the algorithm with some detail. First, the GAP uses the following initial estimate:

$$\hat{\mathbf{x}}_0 = \arg\min_{\mathbf{x}} \|\boldsymbol{\Omega}\mathbf{x}\|_2^2 \quad \text{subject to} \quad \mathbf{y} = \mathbf{M}\mathbf{x}. \tag{18}$$

Not knowing the locations of the cosupport but knowing that many entries of $\boldsymbol{\Omega}\mathbf{x}_0$ are zero, this is a reasonable first estimate of $\mathbf{x}_0$. Once we have $\hat{\mathbf{x}}_0$, we can view $\boldsymbol{\Omega}\hat{\mathbf{x}}_0$ as an estimate of $\boldsymbol{\Omega}\mathbf{x}_0$. Hence, we find the location of the largest entries (in absolute value) of $\boldsymbol{\Omega}\hat{\mathbf{x}}_0$ and regard them as not belonging to the cosupport. After this, we remove the corresponding rows from $\boldsymbol{\Omega}$ and work with a reduced $\boldsymbol{\Omega}$. A detailed description of the algorithm is given in Fig. 5.

Some readers may notice that the GAP has similar flavors to the FOCUSS [26] and the IRLS [12]. This is certainly true in the sense that the GAP solves constrained least squares problems and adjusts weights as it iterates. However, the weight adjustment in the GAP is more aggressive (removal of rows) and binary in nature. We also note that the use of the selection factor t in the GAP is related to Weak Greedy Algorithms [53] for the synthesis model.

Stopping criterion/targeted sparsity In GAP, we have a range of choices between using the full ℓ zeros in the product $\boldsymbol{\Omega}\mathbf{x}$ versus a minimal and sufficient set of $d - m$ zeros. In between these two values, and assuming that the proper elements of Λ have been detected, we expect the solution obtained by the algorithms to be the same, with a slightly better numerical stability for a larger number of zeros.

Thus, an alternative stopping criterion for the GAP could be to detect whether the solution is static or the analysis coefficients of the solution are small. This way, even if the GAP made an error and removed from $\hat{\Lambda}_k$ an index that belongs to the true cosupport Λ, the tendency of the solution to stabilize could help in preventing the algorithm to incorporate this error into the solution. In fact, this criterion is used in the experiment in Section 6.

Multiple selections The selection factor $0 < t \leqslant 1$ allow the selection of multiple rows at once, to accelerate the algorithm by reducing the number of iterations.

Solving the required least squares problems The solution of Eq. (18) (and of the adjusted problems with reduced $\boldsymbol{\Omega}$ at subsequent steps of the algorithm)—under some suitable conditions—is given analytically (see Appendix E for the derivation) by

$$\hat{\mathbf{x}}_0 = \left(\mathbf{M}^T\mathbf{M} + \left(\mathbf{Id} - \mathbf{M}^T\left(\mathbf{M}\mathbf{M}^T\right)\mathbf{M}\right)\boldsymbol{\Omega}^T\boldsymbol{\Omega}\right)^{-1}\mathbf{M}^T\mathbf{y}. \tag{19}$$

- **Task:** Approximate the solution of (15).
- **Parameters:** Given are the matrices $\mathbf{M}$, $\boldsymbol{\Omega}$, the vector $\mathbf{y}$, the target number of zeros ℓ, and a selection factor $t \in (0, 1]$.
- **Initialization:** Set $k = 0$ and perform the following steps:
  - **Initialize Cosupport:** $\hat{\Lambda}_0 = \{1, 2, 3, \ \ldots \ , p\}$,
  - **Initialize Solution:**

$$\hat{\mathbf{x}}_0 = \arg\min_{\mathbf{x}} \|\boldsymbol{\Omega}_{\hat{\Lambda}_0}\mathbf{x}\|_2^2 \quad \text{subject to} \quad \mathbf{y} = \mathbf{M}\mathbf{x}.$$

- **GAP Iterations:** Increment k by 1 and perform the following steps:
  - **Project:** Compute $\alpha = \boldsymbol{\Omega}\hat{\mathbf{x}}_{k-1}$,
  - **Find largest entries:** $\Gamma_k = \{i : \ |\alpha_i| \geqslant t \max_j |\alpha_j|\}$,
  - **Update Support:** $\hat{\Lambda}_k = \hat{\Lambda}_{k-1} \setminus \Gamma_k$, and
  - **Update Solution:**

$$\hat{\mathbf{x}}_k = \arg\min_{\mathbf{x}} \|\boldsymbol{\Omega}_{\hat{\Lambda}_k}\mathbf{x}\|_2^2 \quad \text{subject to} \quad \mathbf{y} = \mathbf{M}\mathbf{x}.$$

  - **Stopping Criterion:** If $k \geqslant p - d + m$ (or $k \geqslant p - \ell$), stop.
- **Output:** The proposed solution is $\hat{\mathbf{x}}_{\mathrm{GAP}} = \hat{\mathbf{x}}_k$ obtained after k iterations.

Fig. 5. Greedy Analysis Pursuit (GAP) algorithm.

In practice, instead of (18), we compute

$$\hat{\mathbf{x}}_0 = \arg\min_{\mathbf{x}} \left\{ \|\mathbf{y} - \mathbf{M}\mathbf{x}\|_2^2 + \lambda \|\boldsymbol{\Omega}\mathbf{x}\|_2^2 \right\} = \arg\min_{\mathbf{x}} \left\| \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{M} \\ \sqrt{\lambda}\boldsymbol{\Omega} \end{bmatrix} \mathbf{x} \right\|_2^2$$

for a small $\lambda > 0$, yielding the solution

$$\hat{\mathbf{x}}_0 = \begin{bmatrix} \mathbf{M} \\ \sqrt{\lambda}\boldsymbol{\Omega} \end{bmatrix}^{\dagger} \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} = \left(\mathbf{M}^T\mathbf{M} + \lambda \boldsymbol{\Omega}^T\boldsymbol{\Omega}\right)^{-1}\mathbf{M}^T\mathbf{y}.$$

We point out that the use of the parameter λ is for convenience in this work but it becomes more useful when dealing with noisy observation. Furthermore, the small value of λ can have adverse effects on computational cost for some numerical algorithms (e.g., the conjugate gradient method) to solve the above minimization. In our implementation, we have used λ values roughly from 10^{-4} to 10^{-6}.

5. Theoretical analysis

So far, we have introduced the cosparse analysis data model, provided uniqueness results in the context of linear inverse problems for the model, and described some algorithms that may be used to solve such linear inverse problems to recover cosparse signals. Before validating the algorithms and the model proposed with experimental results, we first investigate theoretically under what conditions the proposed algorithms to solve cosparse signal recovery (15) are guaranteed to work. After that, we discuss the nature of the condition derived by contrasting it to that for the synthesis model. Further discussion including some desirable properties of $\boldsymbol{\Omega}$ and $\mathbf{M}$ can be found in Appendix D.

5.1. A sufficient condition for the success of the ℓ_1-minimization

In the sparse synthesis framework, there is a well-known necessary and sufficient condition called the *null space property* (NSP) [16] that guarantees the success of the synthesis ℓ_1-minimization

$$\hat{\mathbf{z}}_0 := \operatorname*{argmin}_{\mathbf{z}} \|\mathbf{z}\|_1 \quad \text{subject to} \quad \mathbf{y} = \boldsymbol{\Phi}\mathbf{z} \tag{20}$$

to recover the sparsest solution, say $\mathbf{z}_0$, to $\mathbf{y} = \boldsymbol{\Phi}\mathbf{z}$. To elaborate, in the case of a fixed support T, the ℓ_1-minimization (20) recovers every sparse coefficient vector $\mathbf{z}_0$ supported on T if and only if

$$\|\mathbf{z}_T\|_1 < \|\mathbf{z}_{T^c}\|_1, \quad \forall \mathbf{z} \in \mathrm{Null}(\boldsymbol{\Phi}), \ \mathbf{z} \neq 0. \tag{21}$$

The NSP (21) cannot easily be checked but some 'simpler' sufficient conditions can be derived from it; for example, one can get a recovery condition of [54] called the Exact Recovery Condition (ERC):

$$\left\| \boldsymbol{\Phi}_T^{\dagger} \boldsymbol{\Phi}_{T^c} \right\|_{1 \to 1} < 1, \tag{22}$$

where the notation $|||\mathbf{A}|||_{p\to q}$ denotes the operator norm of $\mathbf{A}$ from ℓ_p to ℓ_q, i.e.,

$$|||\mathbf{A}|||_{p\to q} := \sup_{\mathbf{x}\neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_q}{\|\mathbf{x}\|_p}.$$

The ERC (22) also implies the success of greedy algorithms such as OMP [54]. Note that here we used the symbol $\boldsymbol{\Phi}$ for an object which may be viewed as a dictionary or a measurement matrix. Separating the data model and sampling, we can write $\boldsymbol{\Phi} = \mathbf{MD}$ as was done in Section 3.

One may naturally wonder: is there a condition for the cosparse analysis model that is similar to (21) and (22)? The answer to this question seems to be affirmative with some qualification as the following two results show (the proofs are in Appendix A):

Theorem 7. *Let Λ be a fixed cosupport. The analysis ℓ_1-minimization*

$$\hat{\mathbf{x}}_0 := \operatorname*{argmin}_{\mathbf{x}} \|\boldsymbol{\Omega}\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{y} := \mathbf{M}\mathbf{x}_0 = \mathbf{M}\mathbf{x} \tag{23}$$

recovers every $\mathbf{x}_0$ with cosupport Λ as a unique minimizer if, and only if,

$$\sup_{\mathbf{x}_\Lambda:\, \boldsymbol{\Omega}_\Lambda \mathbf{x}_\Lambda = 0} \left|\left\langle \boldsymbol{\Omega}_{\Lambda^c}\mathbf{z}, \operatorname{sign}(\boldsymbol{\Omega}_{\Lambda^c}\mathbf{x}_\Lambda)\right\rangle\right| < \left\|\boldsymbol{\Omega}_\Lambda \mathbf{z}\right\|_1, \quad \forall \mathbf{z} \in \operatorname{Null}(\mathbf{M}),\ \mathbf{z}\neq 0. \tag{24}$$

Corollary 8. *Let $\mathbf{N}^T$ be any $d\times(d-m)$ basis matrix for the null space* Null($\mathbf{M}$)*, and Λ be a fixed cosupport such that the $\ell\times(d-m)$ matrix $\boldsymbol{\Omega}_\Lambda\mathbf{N}^T$ is of full rank $d-m$. If*

$$\sup_{\mathbf{x}_\Lambda:\, \boldsymbol{\Omega}_\Lambda \mathbf{x}_\Lambda = 0} \left\|\left(\mathbf{N}\boldsymbol{\Omega}_\Lambda^T\right)^\dagger \mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T \operatorname{sign}(\boldsymbol{\Omega}_{\Lambda^c}\mathbf{x}_\Lambda)\right\|_\infty < 1, \tag{25}$$

then the analysis ℓ_1-minimization (23) recovers every $\mathbf{x}_0$ with cosupport Λ. Moreover, if

$$\left|\left|\left|\left(\mathbf{N}\boldsymbol{\Omega}_\Lambda^T\right)^\dagger \mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T\right|\right|\right|_{\infty\to\infty} = \left|\left|\left|\boldsymbol{\Omega}_{\Lambda^c}\mathbf{N}^T\left(\boldsymbol{\Omega}_\Lambda\mathbf{N}^T\right)^\dagger\right|\right|\right|_{1\to 1} < 1 \tag{26}$$

then condition (25) holds true.

There is an apparent similarity between the analysis ERC condition (26) above and its standard synthesis counterpart (22), yet there are some subtle differences between the two that will be highlighted in Section 5.3.

5.2. A sufficient condition for the success of the GAP

There is an interesting parallel between the synthesis ERC (22) and its analysis version in Corollary 8; namely, the analysis ERC condition (26) also implies the success of the GAP algorithm when the selection factor t of the GAP is 1 (in fact, $|||\boldsymbol{\Omega}_{\Lambda^c}\mathbf{N}^T(\boldsymbol{\Omega}_\Lambda\mathbf{N}^T)^\dagger|||_{1\to 1} < t \leqslant 1$), as we will now show.

From the way GAP algorithm works, we can guarantee that it will perform a correct elimination at the first step if the largest analysis coefficients of $\boldsymbol{\Omega}_{\Lambda^c}\hat{\mathbf{x}}_0$ of the first estimate $\hat{\mathbf{x}}_0$ are larger than the largest of $\boldsymbol{\Omega}_\Lambda\hat{\mathbf{x}}_0$ where Λ denotes the true cosupport of $\mathbf{x}_0$. This observation suggests that we can hope to find a condition for success if we can find some relation between $\boldsymbol{\Omega}_{\Lambda^c}\hat{\mathbf{x}}_0$ and $\boldsymbol{\Omega}_\Lambda\hat{\mathbf{x}}_0$. The following result provides such a relation:

Lemma 9. *Let $\mathbf{N}^T$ be any $d\times(d-m)$ basis matrix for the null space* Null($\mathbf{M}$) *and Λ be a fixed cosupport such that the $\ell\times(d-m)$ matrix $\boldsymbol{\Omega}_\Lambda\mathbf{N}^T$ is of full rank $d-m$. Let a signal $\mathbf{x}_0$ with $\boldsymbol{\Omega}_\Lambda\mathbf{x}_0 = 0$ and its observation $\mathbf{y} = \mathbf{M}\mathbf{x}_0$ be given. Then the estimate $\hat{\mathbf{x}}_0$ in (18) satisfies*

$$\boldsymbol{\Omega}_\Lambda\hat{\mathbf{x}}_0 = -\left(\mathbf{N}\boldsymbol{\Omega}_\Lambda^T\right)^\dagger \mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T \boldsymbol{\Omega}_{\Lambda^c}\hat{\mathbf{x}}_0. \tag{27}$$

Having obtained a relation between $\boldsymbol{\Omega}_\Lambda\hat{\mathbf{x}}_0$ and $\boldsymbol{\Omega}_{\Lambda^c}\hat{\mathbf{x}}_0$, we can derive a sufficient condition which guarantees the success of GAP for recovering the true target signal $\mathbf{x}_0$:

Theorem 10. *Let $\mathbf{N}^T$ be any $d\times(d-m)$ basis matrix for the null space* Null($\mathbf{M}$) *and Λ be a fixed cosupport such that the $\ell\times(d-m)$ matrix $\boldsymbol{\Omega}_\Lambda\mathbf{N}^T$ is of full rank $d-m$. Let a signal $\mathbf{x}_0$ with $\boldsymbol{\Omega}_\Lambda\mathbf{x}_0 = 0$ and an observation $\mathbf{y} = \mathbf{M}\mathbf{x}_0$ be given. Suppose that the analysis ERC (26) holds true. Then, when applied to solve (15), GAP with selections factor $t > |||(\mathbf{N}\boldsymbol{\Omega}_\Lambda^T)^\dagger\mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T|||_{\infty\to\infty}$ will recover $\mathbf{x}_0$ after at most $|\Lambda^c|$ iterations.*

Proof. At the first iteration, GAP is doing the correct thing if it removes a row from $\boldsymbol{\Omega}_{\Lambda^c}$. Clearly, this happens when

$$\|\boldsymbol{\Omega}_{\Lambda}\hat{\mathbf{x}}_0\|_\infty < t\|\boldsymbol{\Omega}_{\Lambda^c}\hat{\mathbf{x}}_0\|_\infty. \tag{28}$$

In view of (27), if (26) holds and $t > |||(\mathbf{N}\boldsymbol{\Omega}_{\Lambda}^T)^\dagger \mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T|||_{\infty\to\infty}$, then (28) is guaranteed. Therefore, GAP successfully removes a row from $\boldsymbol{\Omega}_{\Lambda^c}$ at the first step.

Now suppose that (26) was true and GAP has removed a row from $\boldsymbol{\Omega}_{\Lambda^c}$ at the first iteration. Then, at the next iteration, we have the same $\boldsymbol{\Omega}_{\Lambda}$ and, in the place of $\boldsymbol{\Omega}_{\Lambda^c}$, a submatrix $\tilde{\boldsymbol{\Omega}}_{\Lambda^c}$ of $\boldsymbol{\Omega}_{\Lambda^c}$ (with one fewer row). Thus, we can invoke Lemma 9 again and we have

$$\boldsymbol{\Omega}_{\Lambda}\hat{\mathbf{x}}_1 = -\left(\mathbf{N}\boldsymbol{\Omega}_{\Lambda}^T\right)^\dagger \mathbf{N}\tilde{\boldsymbol{\Omega}}_{\Lambda^c}^T \tilde{\boldsymbol{\Omega}}_{\Lambda^c}\hat{\mathbf{x}}_1.$$

Let $\mathbf{R}_0 := (\mathbf{N}\boldsymbol{\Omega}_{\Lambda}^T)^\dagger \mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T$ and $\mathbf{R}_1 := (\mathbf{N}\boldsymbol{\Omega}_{\Lambda}^T)^\dagger \mathbf{N}\tilde{\boldsymbol{\Omega}}_{\Lambda^c}^T$. We observe that $\mathbf{R}_1$ is a submatrix of $\mathbf{R}_0$ obtained by removing one column. Therefore,

$$|||\mathbf{R}_1|||_{\infty\to\infty} \leqslant |||\mathbf{R}_0|||_{\infty\to\infty} < t.$$

By the same logic as for the first step, the success of the second step is guaranteed. Repeating the same argument for the subsequent steps, we obtain the conclusion.

Clearly, at least one row from Λ^c is removed at each iteration. Therefore, $\mathbf{x}_0$ is recovered after at most $|\Lambda^c|$ iterations. □

Remark 11. As pointed out at the beginning of the subsection, the Exact Recovery Condition (26) for the cosparse signal recovery guarantees the success of both the GAP and the analysis ℓ_1-minimization.

5.3. Analysis vs. synthesis exact recovery conditions

When $\boldsymbol{\Phi}$ is written as $\mathbf{MD}$, the exact recovery condition (22) for the sparse synthesis model is equivalent to

$$\left|\left|\left|(\mathbf{MD}_T)^\dagger \mathbf{MD}_{T^c}\right|\right|\right|_{1\to 1} < 1. \tag{29}$$

Here, T is the support of the sparsest representation of the target signal. At first glance, the two conditions (29) and (26):

$$\left|\left|\left|\boldsymbol{\Omega}_{\Lambda^c}\mathbf{N}^T\left(\boldsymbol{\Omega}_{\Lambda}\mathbf{N}^T\right)^\dagger\right|\right|\right|_{1\to 1} < 1$$

look similar; that is, for both cases, one needs to understand the characteristics of a single matrix, $\boldsymbol{\Omega}\mathbf{N}^T$ for the cosparse model, and $\mathbf{MD}$ for the sparse model. Moreover, the expressions involving these matrices have similar forms.

However, upon closer inspection, there is a crucial difference in the structures of the two expressions. In the synthesis case, the operator norm in question depends only on how the *columns* of $\mathbf{MD}$ are related, since a more explicit writing of the pseudo-inverse shows that the matrix to consider is

$$\left(\mathbf{D}_T^T\mathbf{M}^T\mathbf{MD}_T\right)^{-1}(\mathbf{MD}_T)^T\mathbf{MD}_{T^c}.$$

This fact allows us to obtain more easily characterizable conditions like incoherence assumptions [54] that ensure condition (29).

To the contrary, in the analysis case, more complicated relations among *the rows and the columns* of $\boldsymbol{\Omega}\mathbf{N}^T$ have to be taken into account. The matrix to consider being

$$\boldsymbol{\Omega}_{\Lambda^c}\mathbf{N}^T\left(\mathbf{N}\boldsymbol{\Omega}_{\Lambda}^T\boldsymbol{\Omega}_{\Lambda}\mathbf{N}^T\right)^{-1}\mathbf{N}\boldsymbol{\Omega}_{\Lambda}^T,$$

the inner expression $\mathbf{N}\boldsymbol{\Omega}_{\Lambda}^T\boldsymbol{\Omega}_{\Lambda}\mathbf{N}^T$ is connected with how the *columns* of $\boldsymbol{\Omega}\mathbf{N}^T$ are related. However, because the matrices $\boldsymbol{\Omega}_{\Lambda^c}\mathbf{N}^T$ and $\mathbf{N}\boldsymbol{\Omega}_{\Lambda}^T$ appear outside, it also becomes relevant how the *rows* of $\boldsymbol{\Omega}\mathbf{N}^T$ are related.

There is also an interesting distinction in terms of the sharpness of these exact recovery conditions. Namely, the violation of (29) implies the failure of the OMP in the sense that there exist a sparse vector $\mathbf{x} = \mathbf{D}_T\mathbf{z}_T$ for which the first step of OMP picks up an atom which is not indexed by T. To the opposite, the violation of (26) does not seem to imply the necessary "failure" of GAP in a similar sense.

Note however that both conditions are not essential for the success of the algorithms. One of the reasons is that the violation of the conditions does not guarantee that the algorithms would select wrong atoms. Furthermore, even if the GAP or the OMP "fails" in one step, that does not necessarily mean that the algorithms fail in the end: further steps may still enable them to achieve an accurate estimate of the vector $\mathbf{x}_0$.

5.4. Relation to the work by Candès et al. [7]

Before moving on to experimental results, we discuss the recovery guarantee result of Candès et al. [7] for the algorithm

$$\hat{\mathbf{x}} = \underset{\tilde{\mathbf{x}}\in\mathbb{R}^d}{\operatorname{argmin}} \left\|\mathbf{D}^T\tilde{\mathbf{x}}\right\|_1 \quad \text{subject to} \quad \|\mathbf{M}\tilde{\mathbf{x}} - \mathbf{y}\|_2 \leqslant \epsilon, \tag{30}$$

when partial noisy observation $\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{w}$ with $\|\mathbf{w}\|_2 \leqslant \epsilon$ is given for an unknown target signal $\mathbf{x}$.

In order to derive the result, the concept of D-RIP is introduced [7]: A measurement matrix $\mathbf{M}$ satisfies D-RIP adapted to $\mathbf{D}$ with constant δ_s^D if

$$\left(1-\delta_s^D\right)\|v\|_2^2 \leqslant \|\mathbf{M}v\|_2^2 \leqslant \left(1+\delta_s^D\right)\|v\|_2^2$$

holds for all v that can be expressed as a linear combination of s columns of $\mathbf{D}$. With this definition of D-RIP, the main result of [7] can be stated as follows: for an arbitrary tight frame $\mathbf{D}$ and a measurement matrix $\mathbf{M}$ satisfying D-RIP with $\delta_{7s}^D < 0.6$, the solution $\hat{\mathbf{x}}$ to (30) satisfies

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leqslant C_0\epsilon + C_1\frac{\|\mathbf{D}^T\mathbf{x} - (\mathbf{D}^T\mathbf{x})_s\|_1}{\sqrt{s}}, \tag{31}$$

where the constants C_0 and C_1 may depend only on δ_{7s}^D, and the notation $(c)_s$ represents a sequence obtained from a sequence c by keeping the s-largest values of c in magnitude (and setting the others to zero).

The above recovery guarantee is one of the few—very likely the only—results existing in the literature on (30). However, we observe that there is much room for improving the result. We now discuss why we hold this view. For clarity and for the purpose of comparison to our result, we consider only the case $\epsilon = 0$ for (30).

First, we note that [7] implicitly uses the estimate of type $\|\boldsymbol{\Omega}_{\Lambda^c}\mathbf{z}\|_1 < \|\boldsymbol{\Omega}_{\Lambda}\mathbf{z}\|_1$ for (24). Hence, the main result of [7] cannot be sharp in general due to the fact that the sign patterns of (24) are ignored.[5]

Second, the quality of the bound $\|\mathbf{D}^T\mathbf{x} - (\mathbf{D}^T\mathbf{x})_s\|_1/\sqrt{s}$ in (31) is measured in terms of how effective $\mathbf{D}^T\mathbf{x}$ is in sparsifying the signal $\mathbf{x}$ with respect to the dictionary $\mathbf{D}$. To explain, let us consider the synthesis ℓ_1-minimization

$$\Delta_1(\mathbf{x}) := \underset{\mathbf{z}\in\mathbb{R}^n}{\operatorname{argmin}} \|\mathbf{z}\|_1 \quad \text{subject to} \quad \mathbf{M}\mathbf{D}\mathbf{z} = \mathbf{M}\mathbf{x} \tag{32}$$

and let $\Delta_0(\mathbf{x})$ be the sparsest representation of $\mathbf{x}$ with $\mathbf{D}$:

$$\Delta_0(\mathbf{x}) := \underset{\mathbf{z}\in\mathbb{R}^n}{\operatorname{argmin}} \|\mathbf{z}\|_0 \quad \text{subject to} \quad \mathbf{D}\mathbf{z} = \mathbf{x}.$$

Applying the standard result for the synthesis ℓ_1-minimization, we have

$$\left\|\Delta_1(\mathbf{x}) - \Delta_0(\mathbf{x})\right\|_2 \leqslant C_2\frac{\|\Delta_0(\mathbf{x}) - (\Delta_0(\mathbf{x}))_s\|_1}{\sqrt{s}}$$

provided that $\mathbf{M}\mathbf{D}$ satisfies the standard RIP with, e.g., $\delta_{2s} < \sqrt{2}-1 \approx 0.414$. Since $\mathbf{D}$ is a tight frame, it implies

$$\left\|\mathbf{D}\Delta_1(\mathbf{x}) - \mathbf{x}\right\|_2 \leqslant C_2\frac{\|\Delta_0(\mathbf{x}) - (\Delta_0(\mathbf{x}))_s\|_1}{\sqrt{s}}. \tag{33}$$

Note that both $\Delta_0(\mathbf{x})$ and $\mathbf{D}^T\mathbf{x}$ are legitimate representations of $\mathbf{x}$ since $\mathbf{D}\Delta_0(\mathbf{x}) = \mathbf{x} = \mathbf{D}\mathbf{D}^T\mathbf{x}$. Thus, $\Delta_0(\mathbf{x})$ is sparser than $\mathbf{D}^T\mathbf{x}$ in general; in this sense, $\mathbf{D}^T\mathbf{x}$ is not effective in sparsifying $\mathbf{x}$. Given this, we expect that $\|\Delta_0(\mathbf{x}) - (\Delta_0(\mathbf{x}))_s\|_1/\sqrt{s}$ is smaller than $\|\mathbf{D}^T\mathbf{x} - (\mathbf{D}^T\mathbf{x})_s\|_1/\sqrt{s}$. We now see that (31) with $\epsilon = 0$ and (33) are of the same form. Furthermore, given the degree of restriction on the RIP constants ($\delta_{7s}^D < 0.6$ *vs.* $\delta_{2s} < 0.414$), we can only expect that the constant C_2 is smaller than C_1. From these considerations, (31) suggests to us that analysis ℓ_1-minimization (17) performs on par with synthesis ℓ_1-minimization (32), or tends to perform worse.

Third, the only way for (31) to explain that the cosparse signals are perfectly recovered by analysis ℓ_1-minimization is to show that $\mathbf{D}^T\mathbf{x}$ is exactly s-sparse for some $s > 0$ with D-RIP constant $\delta_{7s}^D < 0.6$. Unfortunately, we can quickly observe that the situation becomes hopeless even for moderately overcomplete $\mathbf{D}$; for example, let $\mathbf{D}$ be a 1.15-times overcomplete random tight frame for $\mathbb{R}^d$ and consider recovering $(d-1)$-cosparse signals for the operator $\mathbf{D}^T$. Note that $(d-1)$-cosparse signals $\mathbf{x}$ lead to $(0.15d+1)$-sparse representation $\mathbf{D}^T\mathbf{x}$. This means that we need $\delta_{7(0.15d+1)}^D = \delta_{1.05d+7}^D$ to be smaller than 0.6 to show that $\mathbf{x}$ can be recovered with analysis ℓ_1, which of course cannot happen since $\delta_d^D \geqslant 1$ (unless every element $\mathbf{x}$ in the span of $\mathbf{D}$ is uniquely characterized by its projection $\mathbf{M}\mathbf{x}$, an uninteresting situation that can only occur if either $\mathbf{M}$ is

[5] Note that the same lack of sharpness holds true for our results based on (26), yet we will see that these can actually provide cosparse signal recovery guarantees in simple but nontrivial cases.

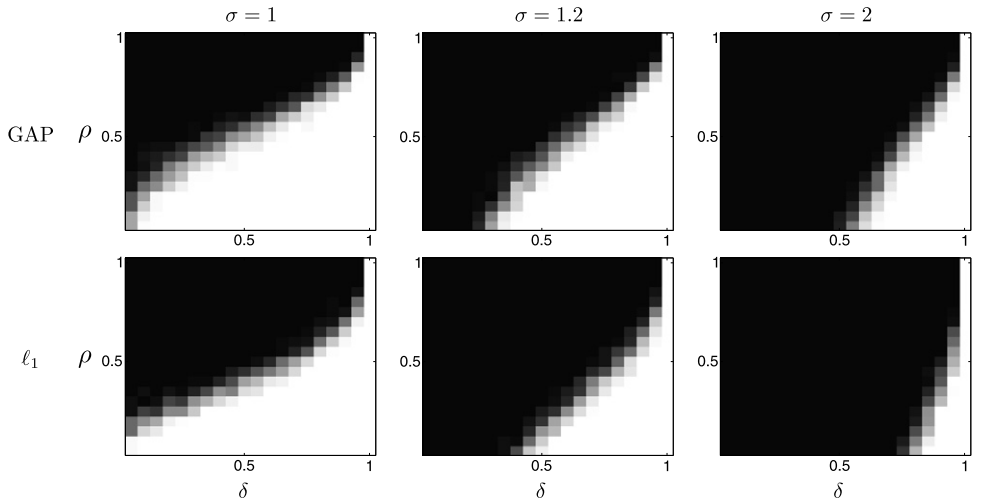

Fig. 6. Recovery rate of analysis algorithms for $d = 200$.

one to one, or $\mathbf{D}$ does not span the signal space). By somehow taking the synthesis view of the signals, (31) cannot explain the recovery of the simplest cosparse signals (cosparsity $d-1$) no matter what $\mathbf{M}$ is (as long as it is underdetermined).

We also observe that the result of [7] cannot say much about the recovery of cosparse signals with respect to the finite difference operators $\boldsymbol{\Omega}_{\mathrm{DIF}}$ discussed in Section 3. This is due to the fact that $\boldsymbol{\Omega}_{\mathrm{DIF}}^T$ is not a tight frame. How does our recovery result (26) fare in this regard? For illustration, we took Ω to be the finite difference operator $\boldsymbol{\Omega}_{\mathrm{DIF}}$ for 32×32 images (thus, $d = 1024$). As a test image, we took $\mathbf{x}$ to be constant in the region $\{(i,j):\ i,j = 1,\dots,16\}$ and $\{(i,j):\ i,j = 1,\dots,16\}^c$. For this admittedly simple test image, using the same notational convention as in Section 3.4, we computed the operator norm in (26) for random measurement matrices $\mathbf{M} \in \mathbb{R}^{640\times 1024}$, with $\mathbf{N}$ a basis of the null space of $\mathbf{M}$ (computed with an SVD), and Λ the cosupport of the test image. When the operator norm was computed for 100 instances $\mathbf{M}$, it was observed to be less than 0.726. Hence, our result does give the guarantee of cosparse signal recovery in simple cases.

6. Experiments

Empirical performance of the proposed algorithms is presented in this section. First, we show how the algorithms perform in synthetic cosparse recovery problems. Second, experimental results for an analysis-based compressed sensing are presented.

6.1. Performance of analysis algorithms

In this section, we apply the algorithms described in Section 4 to synthetic cosparse recovery problems. In the experiment, the entries of $\mathbf{M} \in \mathbb{R}^{m\times d}$ were drawn independently from the normal distribution. The analysis operator $\boldsymbol{\Omega} \in \mathbb{R}^{p\times d}$ was constructed so that its transpose is a random tight frame with unit norm columns for $\mathbb{R}^d$—we will simply say that $\boldsymbol{\Omega}$ is a random tight frame in this case.[6] A random (almost) tight frame $\mathbf{B}$ with unit columns was generated starting from a $d \times p$ Gaussian matrix by alternating the following two steps: (1) Singular value decomposition was performed on $\mathbf{B}$ to yield $\mathbf{USV^T} = \mathbf{B}$, and then $\mathbf{S}$ was replaced with a matrix of the form $[\alpha\mathbf{Id}, \mathbf{0}]$ for $\alpha = \sqrt{p/d}$. This gives us a new $\mathbf{B}$ which is a tight frame. (2) The columns of $\mathbf{B}$ were normalized to unit length. Next, the cosparsity ℓ was chosen, and the true or target signal $\mathbf{x}$ was generated randomly as described in Section 2.2. The observation was obtained by $\mathbf{y} = \mathbf{Mx}$.

Matlab `cvx` package [27] with the default solver SeDuMi [52] was used for the analysis ℓ_1. The precision was set to `best`. For the final results, we used the estimate $\hat{\mathbf{x}}$ from the ℓ_1 solver to obtain an estimate of the cosupport—the cosupport estimate was obtained by taking the indices for which the corresponding analysis coefficient is of size less than 10^{-6}—and then using this cosupport and the observation $\mathbf{y}$ to compute the final estimate of $\mathbf{x}$ (this process can be considered as de-biasing.). No visually noticeable changes due to the last de-biasing step were noted in the results.

Fig. 6 shows the results. In all cases, the signal dimension d is set to 200. We then varied the number m of measurements, the cosparsity ℓ of the target signal, and the operator size p according to the following formulae:

$$m = \delta d, \qquad \ell = d - \rho m, \qquad p = \sigma d,$$

[6] One could also construct $\boldsymbol{\Omega}$ by simply drawing the rows of it randomly and independently from $\mathbf{S}^{d-1}$ without the tight frame constraint. We have run the experiment for such operators and observed that the result was similar.

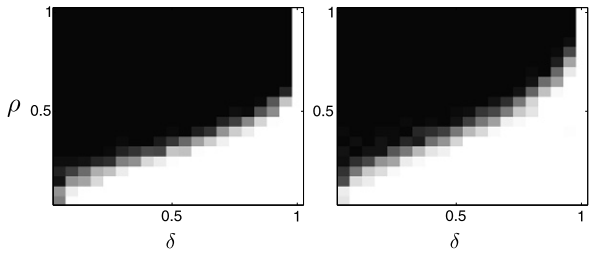

Fig. 7. Recovery rate of analysis algorithms for $d = 200$ and $\sigma = 1$ when $\boldsymbol{\Omega}\mathbf{x}_0$ follows the Rademacher distribution. GAP (left) and L1 (right).

which is consistent with Donoho & Tanner's notations for phase transition diagrams [17]: $\delta = m/d$ is the undersampling ratio, and $\rho = (d-\ell)/m$ measures the relative dimension of the ℓ-cosparse subspaces compared to the number of measures. For every fixed parameter triplet (σ, δ, ρ), the experiment was repeated 50 times. A relative error of size less than 10^{-6} was counted as perfect recovery. Each pixel in the diagrams corresponds to a triplet (σ, δ, ρ) and the pixel intensity represents the ratio of the signals recovered perfectly with white being the 100% success.

The figures show that the GAP can be a viable option when it comes to the cosparse signal recovery. GAP performs better than ℓ_1-minimization, especially for overcomplete $\boldsymbol{\Omega}$'s. It is clear from its description that GAP has polynomial complexity. In practice, computational cost can be high when the size of the problem is very large; to give a rough picture, for the experiment of Section 6.2 (a super greedy version of) GAP was observed to take twice or three times longer to complete the task than `l1 magic`.

It is known that the performance of OMP for the *sparse* signal recovery varies according to the nature of the distribution of the magnitudes of the sparse coefficients. More specifically, OMP performs very well when the coefficients follow independent Gaussian distributions while it does not work as well when the coefficients are drawn from independent Rademacher distributions (1 or -1 with equal probabilities). This phenomenon turns out to be true in the case of GAP as well, and Fig. 7 shows the corresponding result when $\boldsymbol{\Omega}$ is an orthogonal matrix. Note, however, such an unfavorable distribution as Rademacher may not make sense or may have different effects for redundant $\boldsymbol{\Omega}$'s.

An interesting phenomenon observed in the plots for overcomplete $\boldsymbol{\Omega}$ is that there seems to be some threshold δ_* such that if the observation to dimension ratio δ is less than δ_*, one could not recover any signal however cosparse it may be. We may explain this heuristically as follows: If m measurements are available, then the amount of information we have for the signal is $c_1 m$ where c_1 is the number of bits each observation represent. In order to recover a cosparse signal, we need first to identify which subspace the signal belongs to out of $\binom{p}{\ell}$, and then to obtain the $d-\ell$ coefficients for the signal with respect to a basis of the $(d-\ell)$-dimensional subspace. Therefore, roughly speaking, one may hope to recover the signal when

$$c_1 m \geqslant \log_2 \binom{p}{\ell} + c_1(d-\ell) = \log_2 \binom{p}{\ell} + \rho c_1 m.$$

Thus, the recovery is only possible when $(1-\rho)\delta \geqslant \log_2 \binom{p}{d}/(c_1 d)$. Using the relation $p = \sigma d$ and Stirling's approximation, this leads to an asymptotic relation

$$\delta \geqslant (1-\rho)\delta \geqslant \frac{\sigma \log \sigma - (\sigma-1)\log(\sigma-1)}{c_1},$$

which explains the phenomenon.

The calculation above and the experimental evidence from the figures confirm the intuition we had in Section 2.3: the combinatorial number of low-dimensional cosparse subspaces arising from analysis operators in general position is not desirable. This strengthens our view on the necessity of designing/learning analysis operators with high linear dependencies.

6.2. Analysis-based compressed sensing

We observed in Section 6.1 that the cosparse analysis model facilitates effective algorithms to recover partially observed cosparse signals. In this section, we demonstrate the effectiveness of GAP algorithm on a standard toy problem: the Shepp Logan phantom recovery problem.

We consider the following problem that is related to computed tomography (CT): There is an image, say of size $n \times n$, which we are interested in but cannot observe directly. It can only be observed indirectly by means of its 2D Fourier transform coefficients. However, due to high cost of measurements or some physical limitation, the Fourier coefficients can only be observed along a few radial lines. These limited observations or the locations thereof can be modeled by a measurement matrix $\mathbf{M}$, and with the obtained observation we want to recover the original image. As an ideal example, we consider the Shepp Logan phantom. One can easily see that this image is a good example of cosparse signals in $\boldsymbol{\Omega}_{\mathrm{DIF}}$ which consists of all the vertical and horizontal gradients (or one step differences). This image has been used extensively as an example in the literature in the context of compressed sensing (see, e.g., [3,8]).

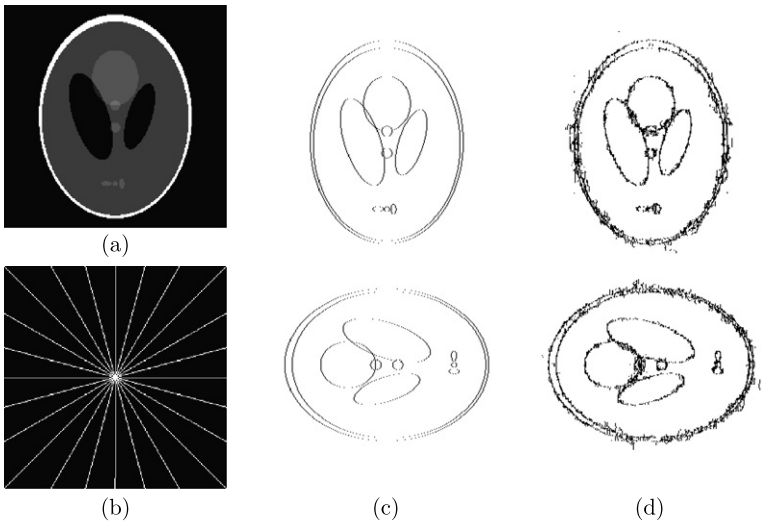

Fig. 8. Recovery of 256×256 Shepp Logan phantom image. (a) Original image. (b) Sampling locations of Fourier coefficients. (c) Locations where one-step difference of the original image is non-zero. Upper half corresponds to the horizontal differences and lower half the vertical differences. (d) Locations that GAP identified/eliminated to be the ones where the differences are likely non-zero. Perfect reconstruction is implied by the fact that this image 'contains' image (c).

Fig. 8 is the result obtained using GAP. The number of measurements that corresponds to 12 radial lines is $m = 3032$. Compared to the number of pixels in the image $d = 65\,536$, it is approximately 4.63%. The number of analysis atoms that give non-zero coefficients is $p - \ell = 2546$. The size of $\boldsymbol{\Omega}_{\mathrm{DIF}}$ is roughly twice the image size $d = 65\,536$, namely $p = 130\,560$. At first glance, this corresponds to very high cosparsity level ($\ell = 130\,560 - 2546$), or put differently, given the high cosparsity level $\ell = 128\,014$, we seem to have required too many measurements. However, using the near optimal guarantee for uniqueness (14), we have a uniqueness guarantee when $m \geqslant 2552$. In view of this, the fact that GAP recovered the signal perfectly for 3032 measurements is encouraging.

We have also ran the GAP algorithm for a larger sized 512×512 problem. The results (not shown here) are visually similar to Fig. 8. In this case, the number of measurements ($m = 7112$) represents approximately 2.71% of the image size ($d = 262\,144$). The number of non-zero analysis coefficients is $p - \ell = 5104$. The sufficient uniqueness condition (14) gives $m \geqslant 5110$ as a number of measurements for the uniqueness.

While encouraged by the result on the phantom images, we acknowledge that these images are unrealistic. Hence, further experiments on more realistic images will be desirable. Some progress has been made [39] to address this.

Remark 12. Due to the large size of these problems, GAP algorithm as described in Section 4 had to be modified: we used numerical optimization (the conjugate gradient method) to approximate pseudo-inverses. Also, due to high computational cost, we eliminated many rows at each iteration (super greedy) instead of one. Although this was not implemented using a selection factor, this can be interpreted as using varying selection factors $0 < t_k < 1$ along the iterations.

To conclude this section, we have repeated the 256×256 Shepp Logan phantom image recovery problem using several algorithms while varying the number of radial observation lines. Given that we know the minimal theoretical number and a theoretically sufficient number of radial observation lines for the uniqueness guarantee, the experimental result gives us an insight on how various algorithms actually perform in the recovery problem in relation to the amount of observation available. Fig. 9 shows the outcome. The algorithms used in the experiment are the GAP, the TV-minimization from `l1magic`, the AIHT from [3], and the back-projection algorithm.[7] The GAP and `l1magic` can be viewed as analysis-based reconstruction algorithms while the AIHT is a synthesis-based reconstruction algorithm. The AIHT is seen to use Haar wavelets as the synthesis dictionary, hence the algorithm implicitly assumes that the phantom image has sparse representation in that dictionary. We remark that while Fig. 9 gives an impression that the AIHT does not have any improvement over the baseline back-projection algorithm, perfect reconstructions were observed for the former when sufficient measurements were available, which is not the case for the back-projection.

Remark 13. It must be noted that in our experiment, each radial line consists of N pixels for an $N \times N$ image; this is in contrast to the fact that the radial lines in the existing codes, e.g. `l1magic`, have $N - 1$ pixels. We have made appropriate

[7] The code for `l1magic` was downloaded from http://www.acm.caltech.edu/l1magic/ and the one for AIHT from http://www.personal.soton.ac.uk/tb1m08/sparsify/AIHT_Paper_Code.zip. The result for the back-projection was obtained using the code for AIHT.

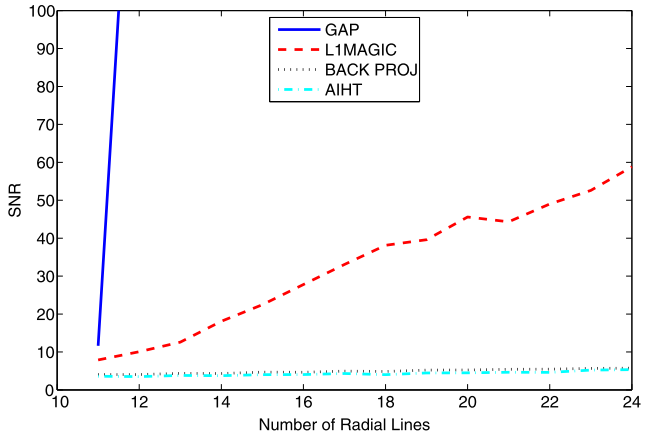

Fig. 9. SNR *vs.* the number of radial observation lines in 256×256 Shepp Logan phantom image recovery. The output line for the GAP is clipped due to high SNR value. SNR was computed as $20\log_{10}(\frac{\|\mathbf{x}\|_2}{\|\hat{\mathbf{x}}-\mathbf{x}\|_2})$ where $\hat{\mathbf{x}}$ is an approximation to the true signal $\mathbf{x}$.

changes for our experiment. The radial lines with $N-1$ pixels do make the recovery problem more difficult and more observations were required for perfect recovery for the GAP.

7. Conclusions and further work

In this work, we have described the cosparse analysis data model as an alternative to the popular sparse synthesis model. We have shown that the cosparse analysis model is distinctly different from the sparse synthesis one in spite of their apparent similarities. In particular, treating the cosparse model as the synthesis model by assuming that the analysis representations of cosparse signals are sparse was demonstrated to be not very meaningful. Having presented the model, we stated conditions that guarantee the uniqueness of cosparse solutions in the context of linear inverse problems based on the work [33]. We then presented some algorithms for the cosparse recovery problem and provided some theoretical result for the analysis ℓ_1-minimization and the newly proposed GAP. Lastly, the model and the proposed algorithm were validated via experimental results.

Although our work in this paper shows that the cosparse analysis model together with algorithms based on the model is an interesting subject to study and viable for practical applications, there is much more to be learned about the model. Among possible future avenues for related research, we list the following[8]:

1. The stability of measurement matrices $\mathbf{M}$ on the analysis union-of-subspaces $\bigcup_{\Lambda} \mathcal{W}_{\Lambda}$.
2. The effect of noise on the cosparse analysis model and associated algorithms.
3. Adaptation to the cases where the signals of interest are not exactly cosparse.
4. The designing/learning of analysis operators for classes of signals of interest.
5. More concrete and/or optimal theoretical success guarantees for algorithms. As an example, one may seek a similar concept as coherence for the analysis operators.
6. Better understanding of the role of linear dependencies between rows of the analysis operator.

On top of these, one can also develop and study more algorithms for cosparse signal recovery; for example, if we view that the GAP is the 'dual' of the OMP, then we could ask: what is the dual of the CoSaMP [41] (or subspace pursuit [11])?

Acknowledgments

We would like to thank the anonymous reviewers for their detailed and insightful comments.

Appendix A. Proof of Theorem 7 and Corollary 8

Let us begin with the simplest case. For a fixed $\mathbf{x}_0$ with cosupport Λ, the analysis ℓ_1-minimization (23) recovers $\mathbf{x}_0$ as the unique minimizer if and only if

$$\left|\left\langle \boldsymbol{\Omega}_{\Lambda^c}\mathbf{z}, \operatorname{sign}(\boldsymbol{\Omega}_{\Lambda^c}\mathbf{x}_0)\right\rangle\right| < \|\boldsymbol{\Omega}_{\Lambda}\mathbf{z}\|_1, \quad \forall \mathbf{z} \in \operatorname{Null}(\mathbf{M}),\ \mathbf{z} \neq 0.$$

[8] Some progress [39] has been made for items 2 and 3 in the listing.

This follows from two facts: (a) the above condition characterizes strict local minima of the optimization problem; (b) the optimization problem is convex and can have at most one strict local minimum, which must be the unique global optimum. From this, we derive the following: The analysis ℓ_1-minimization (23) recovers $\mathbf{x}_0$ as a unique minimizer for *any* $\mathbf{x}_0$ with cosupport Λ, if and only if

$$\sup_{\mathbf{x}_\Lambda:\, \boldsymbol{\Omega}_\Lambda \mathbf{x}_\Lambda = 0} \left|\left\langle \boldsymbol{\Omega}_{\Lambda^c}\mathbf{z}, \operatorname{sign}(\boldsymbol{\Omega}_{\Lambda^c}\mathbf{x}_\Lambda)\right\rangle\right| < \|\boldsymbol{\Omega}_\Lambda \mathbf{z}\|_1, \quad \forall \mathbf{z} \in \operatorname{Null}(\mathbf{M}),\ \mathbf{z} \neq 0$$

and the proof of Theorem 7 is complete.

To obtain Corollary 8, observe that we can remove the constraint $\mathbf{z} \in \operatorname{Null}(\mathbf{M})$ by writing $\mathbf{z} = \mathbf{N}^T\alpha$ where $\mathbf{N}^T$ is an $d \times (d-m)$ basis matrix for $\operatorname{Null}(\mathbf{M})$ and $\alpha \in \mathbb{R}^{d-m}$ is an appropriate coefficient sequence. Thus, the necessary and sufficient condition becomes

$$\sup_{\mathbf{x}_\Lambda:\, \boldsymbol{\Omega}_\Lambda \mathbf{x}_\Lambda = 0} \left|\left\langle \boldsymbol{\Omega}_{\Lambda^c}\mathbf{N}^T\alpha, \operatorname{sign}(\boldsymbol{\Omega}_{\Lambda^c}\mathbf{x}_\Lambda)\right\rangle\right| < \left\|\boldsymbol{\Omega}_\Lambda \mathbf{N}^T\alpha\right\|_1, \quad \forall \alpha \in \mathbb{R}^{d-m},\ \alpha \neq 0. \tag{A.1}$$

Since the $\ell \times (d-m)$ matrix $\boldsymbol{\Omega}_\Lambda\mathbf{N}^T$ is thin ($\ell \geqslant d-m$) and full-rank, defining $\beta := \boldsymbol{\Omega}_\Lambda\mathbf{N}^T\alpha$, we have $\alpha = \left(\boldsymbol{\Omega}_\Lambda\mathbf{N}^T\right)^\dagger\beta$. Therefore, a *sufficient* (but no longer necessary) recovery condition for analysis ℓ_1-minimization is

$$\sup_{\mathbf{x}_\Lambda:\, \boldsymbol{\Omega}_\Lambda \mathbf{x}_\Lambda = 0} \left|\left\langle \boldsymbol{\Omega}_{\Lambda^c}\mathbf{N}^T\left(\boldsymbol{\Omega}_\Lambda\mathbf{N}^T\right)^\dagger\beta, \operatorname{sign}(\boldsymbol{\Omega}_{\Lambda^c}\mathbf{x}_\Lambda)\right\rangle\right| < \|\beta\|_1, \quad \forall \beta \in \mathbb{R}^\ell, \beta \neq 0. \tag{A.2}$$

Equivalently, for all $\mathbf{x}_\Lambda$ with $\boldsymbol{\Omega}_\Lambda\mathbf{x}_\Lambda = 0$,

$$\sup_{\|\beta\|_1 = 1} \left|\left\langle \beta, \left(\mathbf{N}\boldsymbol{\Omega}_\Lambda^T\right)^\dagger\mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T \operatorname{sign}(\boldsymbol{\Omega}_{\Lambda^c}\mathbf{x}_\Lambda)\right\rangle\right| < 1 \tag{A.3}$$

that is to say

$$\sup_{\mathbf{x}_\Lambda:\, \boldsymbol{\Omega}_\Lambda \mathbf{x}_\Lambda = 0} \left\|\left(\mathbf{N}\boldsymbol{\Omega}_\Lambda^T\right)^\dagger\mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T \operatorname{sign}(\boldsymbol{\Omega}_{\Lambda^c}\mathbf{x}_\Lambda)\right\|_\infty < 1. \tag{A.4}$$

Condition (25) follows from the above. To conclude the proof of Corollary 8, we note that since $\|\operatorname{sign}(\boldsymbol{\Omega}_{\Lambda^c}\mathbf{x}_\Lambda)\|_\infty \leqslant 1$, the left-hand side of (A.4) is bounded above by

$$\left\|\left(\mathbf{N}\boldsymbol{\Omega}_\Lambda^T\right)^\dagger\mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T\right\|_{\infty\to\infty} = \left\|\left|\boldsymbol{\Omega}_{\Lambda^c}\mathbf{N}^T\left(\boldsymbol{\Omega}_\Lambda\mathbf{N}^T\right)^\dagger\right|\right\|_{1\to 1}.$$

Therefore, condition (26) implies (25) and the proof is complete.

Appendix B. Proof of Lemma 9

Since $\hat{\mathbf{x}}_0$ is the solution of $\arg\min_{\mathbf{x}} \|\boldsymbol{\Omega}\mathbf{x}\|_2^2$ subject to $\mathbf{y} = \mathbf{M}\mathbf{x}$, applying the Lagrange multiplier method, we observe that $\hat{\mathbf{x}}_0$ satisfies

$$\boldsymbol{\Omega}^T\boldsymbol{\Omega}\hat{\mathbf{x}}_0 = \mathbf{M}^T\mathbf{v} \quad \text{and} \quad \mathbf{M}\hat{\mathbf{x}}_0 = \mathbf{y},$$

for some $\mathbf{v} \in \mathbb{R}^m$. From the first equation, we obtain $\mathbf{v} = (\mathbf{M}^T)^\dagger\boldsymbol{\Omega}^T\boldsymbol{\Omega}\hat{\mathbf{x}}_0$. Putting this back in, one gets $(\mathbf{Id} - \mathbf{M}^T(\mathbf{M}^T)^\dagger) \times \boldsymbol{\Omega}^T\boldsymbol{\Omega}\hat{\mathbf{x}}_0 = 0$. The last equation can be written as $(\mathbf{N}^T)^\dagger\mathbf{N}\boldsymbol{\Omega}^T\boldsymbol{\Omega}\hat{\mathbf{x}}_0 = 0$, where $(\mathbf{N}^T)^\dagger$ is the pseudo-inverse of $\mathbf{N}^T$. Thus,

$$\mathbf{N}\boldsymbol{\Omega}^T\boldsymbol{\Omega}\hat{\mathbf{x}}_0 = 0.$$

Now, we split $\boldsymbol{\Omega}^T\boldsymbol{\Omega} = \boldsymbol{\Omega}_\Lambda^T\boldsymbol{\Omega}_\Lambda + \boldsymbol{\Omega}_{\Lambda^c}^T\boldsymbol{\Omega}_{\Lambda^c}$ and write

$$\mathbf{N}\boldsymbol{\Omega}_\Lambda^T\boldsymbol{\Omega}_\Lambda\hat{\mathbf{x}}_0 = -\mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T\boldsymbol{\Omega}_{\Lambda^c}\hat{\mathbf{x}}_0.$$

Since $\boldsymbol{\Omega}_\Lambda\mathbf{x}_0 = 0$, we can also write

$$\mathbf{N}\boldsymbol{\Omega}_\Lambda^T\boldsymbol{\Omega}_\Lambda\mathbf{u} = -\mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T\boldsymbol{\Omega}_{\Lambda^c}\hat{\mathbf{x}}_0 \tag{B.1}$$

with $\mathbf{u} = \hat{\mathbf{x}}_0 - \mathbf{x}_0$. On the other hand, from $\mathbf{M}\hat{\mathbf{x}}_0 = \mathbf{y} = \mathbf{M}\mathbf{x}_0$, we have $\mathbf{M}\mathbf{u} = 0$. This means that $\mathbf{u}$ can be expressed as $\mathbf{u} =: \mathbf{N}^T\mathbf{w}$ for some $\mathbf{w}$. Plugging this into (B.1), we have

$$\mathbf{N}\boldsymbol{\Omega}_\Lambda^T\boldsymbol{\Omega}_\Lambda\mathbf{N}^T\mathbf{w} = -\mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T\boldsymbol{\Omega}_{\Lambda^c}\hat{\mathbf{x}}_0.$$

Hence, $\mathbf{w} = -(\mathbf{N}\boldsymbol{\Omega}_\Lambda^T\boldsymbol{\Omega}_\Lambda\mathbf{N}^T)^{-1}\mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T\boldsymbol{\Omega}_{\Lambda^c}\hat{\mathbf{x}}_0$. This gives us

$$\hat{\mathbf{x}}_0 - \mathbf{x}_0 = \mathbf{u} = -\mathbf{N}^T\left(\mathbf{N}\boldsymbol{\Omega}_\Lambda^T\boldsymbol{\Omega}_\Lambda\mathbf{N}^T\right)^{-1}\mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T\boldsymbol{\Omega}_{\Lambda^c}\hat{\mathbf{x}}_0.$$

Again, using $\boldsymbol{\Omega}_\Lambda\mathbf{x}_0 = 0$, we have

$$\boldsymbol{\Omega}_\Lambda\hat{\mathbf{x}}_0 = -\boldsymbol{\Omega}_\Lambda\mathbf{N}^T\left(\mathbf{N}\boldsymbol{\Omega}_\Lambda^T\boldsymbol{\Omega}_\Lambda\mathbf{N}^T\right)^{-1}\mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T\boldsymbol{\Omega}_{\Lambda^c}\hat{\mathbf{x}}_0 = -\left(\mathbf{N}\boldsymbol{\Omega}_\Lambda^T\right)^\dagger\mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T\boldsymbol{\Omega}_{\Lambda^c}\hat{\mathbf{x}}_0.$$

Appendix C. Proof of Proposition 6

All the statements in this section are about a 2D regular graph consisting of $d = N \times N$ vertices (V) and the vertical and horizontal edges (E) connecting these vertices. To prove the proposition, we will need three basic lemmas.

Lemma 14. *For a fixed ℓ, the value*

$$\alpha(\ell) := \min_{\Lambda \subset E:\ |\Lambda| \geqslant \ell} \left\{ \left|V(\Lambda)\right| - J(\Lambda) \right\}$$

is achieved for a subgraph $(V(\Lambda), \Lambda)$—we will simply identify Λ with the subgraph from here on—satisfying $|\Lambda| = \ell$ and $J(\Lambda) = 1$.

Proof. It is not difficult to check that the minimum if achieved for Λ with $|\Lambda| = \ell$. Thus, we will assume $|\Lambda| = \ell$.

Now, we need to show that there is also a Λ with $J(\Lambda) = 1$. Suppose that $\tilde{\Lambda}$ with $|\tilde{\Lambda}| = \ell$ achieves $\alpha(\ell)$ and $J(\tilde{\Lambda}) > 1$. We will show that we can obtain Λ from $\tilde{\Lambda}$ that also achieves the value $\alpha(\ell)$, and $|\Lambda| = \ell$ and $J(\Lambda) = 1$. For simplicity, we will consider the case $J(\tilde{\Lambda}) = 2$ only; one can deal with other cases by the repetition of the same argument.

Let $\tilde{\Lambda}_1$ and $\tilde{\Lambda}_2$ be the two connected components of $\tilde{\Lambda}$. Note that on a 2D regular graph, we can shift a subgraph horizontally or vertically unless the subgraph has vertices on all four boundaries of V. Since $\tilde{\Lambda}_1$ and $\tilde{\Lambda}_2$ are disconnected, not all of them can have vertices on all four boundaries of V. Therefore, one of them, say $\tilde{\Lambda}_1$, can be shifted towards the other. Let us consider the first moment when they touched each other. Let t be the number of vertices that coincided. Then, at most $t-1$ edges must have coincided. Thus, denoting the number of edges coincided by $s < t$, the resulting subgraph $\tilde{\Lambda}'$ has $|V(\tilde{\Lambda})| - t$ vertices and $|\tilde{\Lambda}| - s$ edges and one connected components. Now let Λ be a subgraph obtained from $\tilde{\Lambda}'$ by adding s additional edges that are connected to $\tilde{\Lambda}'$. Then,

$$\left|V(\Lambda)\right| \leqslant \left|V\left(\tilde{\Lambda}'\right)\right| + s \leqslant \left|V(\tilde{\Lambda})\right| - t + s,$$

$|\Lambda| = |\tilde{\Lambda}| = \ell$, and $J(\Lambda) = 1$. Hence,

$$\left|V(\Lambda)\right| - J(\Lambda) \leqslant \left|V(\tilde{\Lambda})\right| - t + s - 1 = \left|V(\tilde{\Lambda})\right| - J(\tilde{\Lambda}) - t + s + 1 \leqslant \left|V(\tilde{\Lambda})\right| - J(\tilde{\Lambda}),$$

which is what we wanted to show. □

The next lemma provides a lower bound for the minimum number of vertices $\min_{\Lambda, |\Lambda| = l} |V(\Lambda)|$.

Lemma 15. *For $\ell \geqslant 1$,*

$$\min_{\Lambda, |\Lambda| = \ell} \left|V(\Lambda)\right| \geqslant \frac{\ell}{2} + \frac{1}{2} + \left(\frac{\ell}{2} + \frac{1}{4}\right)^{1/2}.$$

Proof. From Lemma 14 we can restrict ourselves to sets Λ that are connected. Let $e_{\downarrow}(v)$ denote the edge descending from a vertex v, for which we may need to extend the boundary of the lattice. Similarly let $e_{\rightarrow}(v)$ denote the edge extending rightwards from v. We can now define the following *enlargement* of the edge set Λ.

$$\bar{\Lambda} = \left\{ e_{\downarrow}(v), e_{\rightarrow}(v):\ v \in V(\Lambda) \right\}. \tag{C.1}$$

Since each edge can only descend or extend rightwards from a single vertex we have $|\bar{\Lambda}| = 2|V(\Lambda)|$. We also have $\Lambda \subset \bar{\Lambda}$.

We now wish to estimate how much larger $|\bar{\Lambda}|$ is to $|\Lambda|$. Let us define the width, w, of $V(\Lambda)$ as the number of columns spanned by $V(\Lambda)$. Similarly define the height, h, as the number of rows spanned by $V(\Lambda)$. For every row spanned by $V(\Lambda)$ the edge extending rightwards from the right-most vertex is in $\bar{\Lambda} \backslash \Lambda$. Similarly for every column spanned by $V(\Lambda)$ the edge descending from the lowest vertex is also in $\bar{\Lambda} \backslash \Lambda$. Hence we have the following bound

$$2\left|V(\Lambda)\right| = |\bar{\Lambda}| \geqslant |\Lambda| + w + h. \tag{C.2}$$

Intuitively to minimize $|V(\Lambda)|$ we should choose a set of vertices with maximal area to perimeter ratio.

Given that $V(\Lambda)$ lies in a rectangle of size $h \times w$ we can bound the number of edges in Λ using a counting argument to obtain:

$$\ell \leqslant h(w-1) + w(h-1). \tag{C.3}$$

Substituting this into (C.2) we get:

$$2\left|V(\Lambda)\right| \geqslant \ell + w + \frac{\ell + w}{2w - 1} \tag{C.4}$$

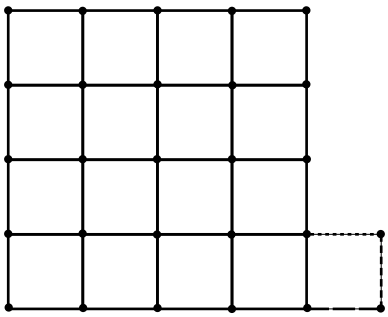

Fig. 10. Add dashed edges (from longer to shorter dashed) to $r \times r$ square subgraph (solid lines).

and similarly we can now bound the minimum possible number of edges by

$$\min_{\Lambda,|\Lambda|=\ell} \big|V(\Lambda)\big| \geqslant \frac{1}{2}\left(\ell + \min_{w\geqslant 1}\left[w + \frac{\ell + w}{2w-1}\right]\right), \tag{C.5}$$

where on the right-hand side we are minimizing over all $w \geqslant 1$. Although this includes non-integer values of w this does not invalidate the bound. It only makes it less tight.

The right-hand side of (C.5) is convex over $w \geqslant 1$ with the minimum occurring at $w^\star = \frac{1}{2} + (\frac{\ell}{2} + \frac{1}{4})^{1/2}$ for which we also have:

$$w^\star = \frac{(\ell + w^\star)}{(2w^\star - 1)},$$

i.e. square cosupports are optimal.

Inserting this into (C.5) then gives:

$$\min_{\Lambda,|\Lambda|=\ell} \big|V(\Lambda)\big| \geqslant \frac{\ell}{2} + \frac{1}{2} + \left(\frac{\ell}{2} + \frac{1}{4}\right)^{1/2} \tag{C.6}$$

as required. □

The goal of our third lemma is not just to derive a lower bound on $\kappa_{\boldsymbol{\Omega}_{\mathrm{DIF}}}$ but a lower bound that is close to optimal. By Lemma 14, $\kappa_{\boldsymbol{\Omega}_{\mathrm{DIF}}}(\ell)$ is achieved for connected Λ, so, as with Lemma 15, we will consider such Λ's only ($J(\Lambda) = 1$). With $J(\Lambda) = 1$, the formula (12) tells us to look for the cases when $|V(\Lambda)|$ is minimal in order to compute $\kappa_{\boldsymbol{\Omega}_{\mathrm{DIF}}}(\ell)$.

What is the shape of the collection of edges Λ yielding the minimum? Recalling Euler's formula for graphs on plane:

$$\big|V(\Lambda)\big| - |\Lambda| + \big|F(\Lambda)\big| = 2, \tag{C.7}$$

where $F(\Lambda)$ is the faces of Λ which includes the 'unbounded one', we see that we are seeking Λ such that $|F(\Lambda)|$ is maximal, i.e., there is maximum number of faces. By intuition, we conjecture that this happens when Λ consists of all the edges in an almost square, by which we mean $V(\Lambda)$ is an $r \times r$ or $r \times (r+1)$ rectangular grid or the inbetweens (e.g., an $r \times r$ grid of pixels to which $1 \leqslant j \leqslant r$ pixels have been added on one side). These considerations lead to the following:

Lemma 16.

$$\alpha(\ell) \leqslant \frac{\ell}{2} + \sqrt{\frac{\ell}{2}} + 1$$

for $\ell \geqslant 5$.

Proof. For $r \geqslant 2$, we consider a subgraph corresponding to an $r \times r$ square (solid lines) and consider graphs obtained by adding additional edges in the fashion depicted in Fig. 10.

We find that for the square Λ, $|\Lambda| = 2(r^2 - r)$ and $|V(\Lambda)| = r^2$, for the graph Λ with one additional edge, $|\Lambda| = 2(r^2 - r) + 1$ and $|V(\Lambda)| = r^2 + 1$, for the graph Λ with two additional edges, $|\Lambda| = 2(r^2 - r) + 2$ and $|V(\Lambda)| = r^2 + 2$, and for the graph Λ with three additional edges, $|\Lambda| = 2(r^2 - r) + 3$ and $|V(\Lambda)| = r^2 + 2$. In fact, we observe that two edges can be added while adding one additional vertex until Λ corresponds to $r \times (r+1)$ rectangle. Summarizing all these, a graph Λ that is constructed as above, is contained $r \times (r+1)$ rectangle (included), and contains $r \times r$ square; satisfies either $|\Lambda| = 2(r^2 - r) + 2j$ or $|\Lambda| = 2(r^2 - r) + 2j + 1$, and $|V(\Lambda)| = r^2 + j + 1$, for $j = 1, \ldots, r-1$—this holds for $j = r$ as well. (Here, the case $|\Lambda| = 2(r^2 - r) + 1$ is not stated.) By a similar observation, we observe that a graph Λ that is constructed similarly as above, is contained in $(r+1) \times (r+1)$ square (included), and contains $r \times (r+1)$ square; satisfies

either $|\Lambda| = 2r^2 - 1 + 2j$ or $|\Lambda| = 2r^2 - 1 + 2j + 1$, and $|V(\Lambda)| = r^2 + r + j + 1$, for $j = 1, \dots, r$—this holds for $j = r + 1$ as well. Of course, in all cases, $J(\Lambda) = 1$.

The above observation leads to the following inequalities—which we conjecture to be in fact equalities:

$$\alpha\big(2\big(r^2 - r\big) + 2j\big) \leqslant r^2 + j, \quad j = 1, \dots, r,$$
$$\alpha\big(2\big(r^2 - r\big) + 2j + 1\big) \leqslant r^2 + j, \quad j = 1, \dots, r,$$
$$\alpha\big(2r^2 - 1 + 2j\big) \leqslant r^2 + r + j, \quad j = 1, \dots, r + 1,$$
$$\alpha\big(2r^2 - 1 + 2j + 1\big) \leqslant r^2 + r + j, \quad j = 1, \dots, r + 1.$$

We will now express these in a simpler form in terms of $|\Lambda| = \ell$. In the first case, letting $\ell = 2(r^2 - r) + 2j$, we have

$$r^2 + j = \frac{\ell}{2} + r.$$

Since

$$2\big(r^2 - 2r + 1\big) \leqslant 2\big(r^2 - r + 1\big) \leqslant \ell \leqslant 2r^2,$$

we have $r - 1 \leqslant \sqrt{\frac{\ell}{2}} \leqslant r$. Hence, we can write $\alpha(\ell) \leqslant \frac{\ell}{2} + \sqrt{\frac{\ell}{2}} + 1$. The other three cases can be treated similarly and we obtain

$$\alpha(\ell) \leqslant \frac{\ell}{2} + \sqrt{\frac{\ell}{2}},$$
$$\alpha(\ell) \leqslant \frac{\ell}{2} + \sqrt{\frac{\ell}{2}} + \frac{1}{2},$$
$$\alpha(\ell) \leqslant \frac{\ell}{2} + \sqrt{\frac{\ell}{2}}.$$

Therefore, for all $\ell \geqslant 5$, we have $\alpha(\ell) \leqslant \frac{\ell}{2} + \sqrt{\frac{\ell}{2}} + 1$. □

We now put these ingredients together.

Proof of Proposition 6. The proof of the lower bound comes directly from Lemma 16.

To prove the upper bound we note that from Lemma 14, Lemma 15 and Eq. (12) we have:

$$\kappa_{\boldsymbol{\Omega}}(\ell) = |V| - \min_{|\Lambda| \geqslant \ell} \big\{ \big|V(\Lambda)\big| - J(\Lambda) \big\} = d - \min_{|\Lambda| \geqslant \ell} \big|V(\Lambda)\big| + 1$$
$$\leqslant d - \frac{\ell}{2} - \frac{1}{2} - \left(\frac{\ell}{2} + \frac{1}{4}\right)^{1/2} + 1 \leqslant d - \frac{\ell}{2} - \sqrt{\frac{\ell}{2}} + \frac{1}{2}$$

as required. □

Appendix D. Discussion on the analysis exact recovery condition

We observe that the analysis ERC condition (26) is not sharp in general, especially for the redundant $\boldsymbol{\Omega}$. In the case of GAP, tracing the arguments of Lemma 9 and Theorem 10, we conclude that in order for (26) to be sharp, there must exist a cosparse signal $\mathbf{x}_0$ such that, with $\hat{\mathbf{x}}_0$ defined as in (18), $\boldsymbol{\Omega}_{\Lambda^c}\hat{\mathbf{x}}_0$ matches the exact sign pattern of the row of $(\mathbf{N}\boldsymbol{\Omega}_{\Lambda}^T)^{\dagger}\mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T$ with the largest ℓ_1-norm *and* is of constant magnitude in absolute value. We remind that $\hat{\mathbf{x}}_0$ is the initial estimate that appears in the algorithm. Since the collection of $\boldsymbol{\Omega}_{\Lambda^c}\hat{\mathbf{x}}_0$ may not span the whole $\mathbb{R}^{\Lambda^c}$, especially when $\boldsymbol{\Omega}$ is overcomplete, it is unreasonable to expect the existence of such an $\mathbf{x}_0$. Similarly, in the case of analysis ℓ_1, we know that (26) is obtained from (25) in a crude way without taking into account the sign patterns of $\boldsymbol{\Omega}_{\Lambda^c}\mathbf{x}_{\Lambda}$, which is not sharp in general for redundant $\boldsymbol{\Omega}$.

D.1. Average case performance guarantees?

Can we think of a way to obtain a more realistic success guarantee? We have a partial answer for this question in the sense that we can derive a condition—which is not a guarantee—that reflects empirical results more faithfully. The idea is, instead of obtaining an upper bound of the left-hand side of (25) by disregarding (or considering the worst case of) sign patterns, to model the effects of the sign patterns by estimating the size of the left-hand side in terms of the maximum ℓ_2-norm of the rows of $(\mathbf{N}\boldsymbol{\Omega}_{\Lambda}^T)^{\dagger}\mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T$ (up to some constants). Though further investigation is desirable, we have empirically observed that the condition derived in this way better reflected the success rates of GAP and ℓ_1-minimization.

D.2. Desirable properties for $\boldsymbol{\Omega}$ and $\mathbf{M}$

At this point, one may ask a practical question: what are desirable properties of $\boldsymbol{\Omega}$ and $\mathbf{M}$ that would help the performance of GAP or ℓ_1-minimization? Can we gain some insights from our theoretical result? For this, we look for scenarios where the entries of $\mathbf{R}_0 := \boldsymbol{\Omega}_\Lambda \mathbf{N}^T (\mathbf{N}\boldsymbol{\Omega}_\Lambda^T \boldsymbol{\Omega}_\Lambda \mathbf{N}^T)^{-1} \mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T$ are small (hence, it is likely that condition (26) is satisfied). We start with the inner expression $(\mathbf{N}\boldsymbol{\Omega}_\Lambda^T \boldsymbol{\Omega}_\Lambda \mathbf{N}^T)^{-1}$. The larger the minimum singular value of $\mathbf{N}\boldsymbol{\Omega}_\Lambda^T \boldsymbol{\Omega}_\Lambda \mathbf{N}^T$, the smaller the entries of $\mathbf{R}_0$. First, assuming that the rows of $\boldsymbol{\Omega}$ are normalized, we note that the minimum singular value is larger when the size Λ is larger. Second, the closer the minimum singular value is to the maximum one (this is in some sense an RIP-like condition for $\boldsymbol{\Omega}$), the larger it is. These two observations tell us that $\boldsymbol{\Omega}$ should have high linear dependencies (to allow large cosupport Λ) and the rows of $\boldsymbol{\Omega}$ should be close to uniformly distributed on $\mathbf{S}^{d-1}$.

Suppose that $\boldsymbol{\Omega}$ has the properties described above. Then, $\mathbf{R}_0$ is well-approximated by $\mathbf{R}_1 := \gamma \boldsymbol{\Omega}_\Lambda \mathbf{N}^T \mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T$ for some $\gamma > 0$. Therefore, we ask when the entries of $\boldsymbol{\Omega}_\Lambda \mathbf{N}^T \mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T$ are small. Each entry of $\boldsymbol{\Omega}_\Lambda \mathbf{N}^T \mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T$ can be guaranteed to be small if $\mathbf{N}$ satisfies an RIP condition for the space spanned by two rows of $\boldsymbol{\Omega}$ and the rows of $\boldsymbol{\Omega}$ are incoherent. In summary, it is desirable that:

- The rows of $\boldsymbol{\Omega}$ are close to uniformly distributed in $\mathbf{S}^{d-1}$.
- $\boldsymbol{\Omega}$ is highly redundant and has highly linearly dependent structure.
- $\mathbf{M}$ is 'independent' from $\boldsymbol{\Omega}$. This has to do with the RIP-like properties.
- The rows of $\boldsymbol{\Omega}$ are incoherent.
- The cosparsity ℓ is large.

Remark 17. The 2D finite difference operator $\boldsymbol{\Omega}_{\mathrm{DIF}}$ may be considered incoherent even though the coherence is relatively large (1/4). This is because the majority of pairs of rows of $\boldsymbol{\Omega}_{\mathrm{DIF}}$ are in fact uncorrelated.

D.3. Heuristic comparison of success guarantees for analysis ℓ_1 and GAP

We point out that one can obtain from (27) a condition for the GAP that is similar to (25). For this, we observe from (27) that

$$\|\boldsymbol{\Omega}_\Lambda \hat{\mathbf{x}}_0\|_\infty = \left\| \left(\mathbf{N}\boldsymbol{\Omega}_\Lambda^T\right)^\dagger \mathbf{N}\boldsymbol{\Omega}_{\Lambda^c}^T \boldsymbol{\Omega}_{\Lambda^c} \hat{\mathbf{x}}_0 \right\|_\infty = \left\| \left[\boldsymbol{\Omega}_{\Lambda^c} \mathbf{N}^T \left(\boldsymbol{\Omega}_\Lambda \mathbf{N}^T\right)^\dagger\right]^T \boldsymbol{\Omega}_{\Lambda^c} \hat{\mathbf{x}}_0 \right\|_\infty.$$

Since $\|\boldsymbol{\Omega}_\Lambda \hat{\mathbf{x}}_0\|_\infty < \|\boldsymbol{\Omega}_{\Lambda^c} \hat{\mathbf{x}}_0\|_\infty$ is the necessary and sufficient condition for the (one-step) success of the GAP, we can derive a necessary and sufficient condition:

$$\left\| \left[\boldsymbol{\Omega}_{\Lambda^c} \mathbf{N}^T \left(\boldsymbol{\Omega}_\Lambda \mathbf{N}^T\right)^\dagger\right]^T \boldsymbol{\Omega}_{\Lambda^c} \hat{\mathbf{x}}_0 \right\|_\infty < \|\boldsymbol{\Omega}_{\Lambda^c} \hat{\mathbf{x}}_0\|_\infty$$

where $\mathbf{x}_0$ is varied over all signals with cosupport Λ and $\hat{\mathbf{x}}_0$ is the signal resulting from the first step of GAP. The above condition can be rewritten in a form similar to (25):

$$\sup_{\mathbf{x}_0} \left\| \left[\boldsymbol{\Omega}_{\Lambda^c} \mathbf{N}^T \left(\boldsymbol{\Omega}_\Lambda \mathbf{N}^T\right)^\dagger\right]^T \left(\operatorname{sign}(\boldsymbol{\Omega}_{\Lambda^c} \hat{\mathbf{x}}_0) \odot \mathbf{v}\right) \right\|_\infty < 1, \tag{D.1}$$

where $\hat{\mathbf{x}}_0$ is derived as in (18), $\odot$ denotes the element-wise multiplication of vectors, and $\mathbf{v}$ is obtained from $\boldsymbol{\Omega}_{\Lambda^c} \hat{\mathbf{x}}_0$ by taking element-wise absolute values and normalizing it to a unit ℓ_∞-norm ($\mathbf{v} := |\boldsymbol{\Omega}_{\Lambda^c} \hat{\mathbf{x}}_0| / \|\boldsymbol{\Omega}_{\Lambda^c} \hat{\mathbf{x}}_0\|_\infty$). Condition (D.1) and (25) are in a similar form, but there are two differences between the two: first, for (D.1), the signal $\hat{\mathbf{x}}_0$ that appears is not in general a vector with cosupport Λ. It is rather a signal that arises from an approximation. Second, there is a 'weight' vector $\mathbf{v}$ in (D.1). One can heuristically deduce that such a $\mathbf{v}$ favors condition (D.1) to hold true since the size of most entries of $\mathbf{v}$ likely be smaller than 1. Beside these differences, one should keep in mind that condition (D.1) is only for one step.

Appendix E. Derivation of the solution (19)

By the method of Lagrange multiplier, $\hat{\mathbf{x}}_0$ is the solution of (18) if there is $\mu \in \mathbb{R}^m$ such that

$$\boldsymbol{\Omega}^T \boldsymbol{\Omega} \hat{\mathbf{x}}_0 = \mathbf{M}^T \mu, \qquad \mathbf{y} = \mathbf{M}\hat{\mathbf{x}}_0.$$

We first assume that $\mathbf{M}$ is of full rank, thus $\mathbf{M}\mathbf{M}^T$ is invertible. With this assumption, we can solve for μ and obtain $\mu = (\mathbf{M}\mathbf{M}^T)^{-1} \mathbf{M}\boldsymbol{\Omega}^T \boldsymbol{\Omega} \hat{\mathbf{x}}_0$. Substituting this in and combining the two equation, we have

$$\begin{bmatrix} (\mathbf{Id} - \mathbf{M}^T (\mathbf{M}\mathbf{M}^T)^{-1} \mathbf{M}) \boldsymbol{\Omega}^T \boldsymbol{\Omega} \\ \mathbf{M} \end{bmatrix} \hat{\mathbf{x}}_0 = \begin{bmatrix} \mathbf{0} \\ \mathbf{y} \end{bmatrix}.$$

Now we assume that the matrix on the left-hand side is of full rank. Multiplying the both sides by

$$\begin{bmatrix} (\mathbf{Id} - \mathbf{M}^T (\mathbf{M}\mathbf{M}^T)^{-1} \mathbf{M}) & \mathbf{M}^T \end{bmatrix}$$

(which is of full rank), we arrive at

$$\left(\mathbf{M}^T\mathbf{M} + \left(\mathbf{Id} - \mathbf{M}^T\left(\mathbf{M}\mathbf{M}^T\right)^{-1}\mathbf{M}\right)\boldsymbol{\Omega}^T\boldsymbol{\Omega}\right)\hat{\mathbf{x}}_0 = \mathbf{M}^T\mathbf{y},$$

from which (19) follows. Above, we have used the fact that $\mathbf{Id} - \mathbf{M}^T(\mathbf{M}\mathbf{M}^T)^{-1}\mathbf{M}$ is a projection.

References

[1] M. Aharon, M. Elad, A.M. Bruckstein, K-SVD: an algorithm for designing of overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process. 54 (2006) 4311–4322.
[2] Z. Ben-Haim, Y.C. Eldar, M. Elad, Coherence-based performance guarantees for estimating a sparse vector under random noise, IEEE Trans. Signal Process. 58 (10) (2010) 5030–5043.
[3] T. Blumensath, Accelerated iterative hard thresholding, preprint, 2011.
[4] T. Blumensath, M.E. Davies, Sampling theorems for signals from the union of finite-dimensional linear subspaces, IEEE Trans. Inform. Theory 55 (4) (2009) 1872–1882.
[5] A.M. Bruckstein, D.L. Donoho, M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, SIAM Rev. 51 (1) (February 2009) 34–81.
[6] J.-F. Cai, S. Osher, Z. Shen, Split Bregman methods and frame based image restoration, Multiscale Model. Simul. 8 (2) (2009) 337–369.
[7] E.J. Candès, Y.C. Eldar, D. Needell, P. Randall, Compressed sensing with coherent and redundant dictionaries, Appl. Comput. Harmon. Anal. 31 (2011) 59–73.
[8] E.J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, IEEE Trans. Inform. Theory 52 (2) (February 2006) 489–509.
[9] E.J. Candès, T. Tao, The Dantzig selector: statistical estimation when p is much larger than n, Ann. Statist. 35 (6) (2007) 2313–2351.
[10] S.S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit, SIAM J. Sci. Comput. 20 (1998) 33–61.
[11] W. Dai, O. Milenkovic, Subspace pursuit for compressive sensing signal reconstruction, IEEE Trans. Inform. Theory 55 (5) (May 2009) 2230–2249.
[12] I. Daubechies, R. DeVore, M. Fornasier, C.S. Güntürk, Iteratively reweighted least squares minimization for sparse recovery, Comm. Pure Appl. Math. 63 (2010) 1–38.
[13] G. Davis, S. Mallat, M. Avellaneda, Adaptive greedy approximations, Constr. Approx. 13 (1) (1997) 57–98.
[14] M.N. Do, M. Vetterli, The contourlet transform: an efficient directional multiresolution image representation, IEEE Trans. Image Process. 14 (2005) 2091–2106.
[15] D.L. Donoho, M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization, Proc. Natl. Acad. Sci. USA 100 (5) (March 2003) 2197–2202.
[16] D.L. Donoho, X. Huo, Uncertainty principles and ideal atomic decomposition, IEEE Trans. Inform. Theory 47 (7) (2001) 2845–2862.
[17] D.L. Donoho, J. Tanner, Counting faces of randomly-projected polytopes when the projection radically lowers dimension, J. Amer. Math. Soc 22 (1) (January 2009) 1–53.
[18] P.L. Dragotti, M. Vetterli, Wavelet footprints: theory, algorithms, and applications, IEEE Trans. Signal Process. 51 (5) (May 2003) 1306–1323.
[19] M. Elad, Sparse and Redundant Representations – From Theory to Applications in Signal and Image Processing, Springer, 2010.
[20] M. Elad, M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries, IEEE Trans. Image Process. 15 (2006) 3736–3745.
[21] M. Elad, P. Milanfar, R. Rubinstein, Analysis versus synthesis in signal priors, Inverse Problems 23 (3) (June 2007) 947–968.
[22] M. Elad, J.-L. Starck, P. Querre, D.L. Donoho, Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA), Appl. Comput. Harmon. Anal. 19 (2005) 340–358.
[23] K. Engan, S.O. Aase, J.H. Husoy, Multi-frame compression: theory and design, Signal Process. 80 (2000) 2121–2140.
[24] S. Farsiu, D. Robinson, M. Elad, P. Milanfar, Advances and challenges in super-resolution, Int. J. Imaging Syst. Technol. 14 (2004) 47–57.
[25] M.R. Garey, D.S. Johnson, Computers and Intractability; A Guide to the Theory of NP-Completeness, W.H. Freeman & Co., New York, NY, USA, 1990.
[26] I.F. Gorodnitsky, B.D. Rao, Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm, IEEE Trans. Signal Process. (1997) 600–616.
[27] M. Grant, S. Boyd, Y. Ye, CVX: Matlab software for disciplined convex programming, August 2008.
[28] R. Gribonval, M. Nielsen, Sparse representations in unions of bases, IEEE Trans. Inform. Theory 49 (12) (December 2003) 3320–3325.
[29] R. Gribonval, M. Nielsen, Highly sparse representations from dictionaries are unique and independent of the sparseness measure, Appl. Comput. Harmon. Anal. 22 (3) (May 2007) 335–355.
[30] S. Kluckner, T. Pock, H. Bischof, Exploiting redundancy for aerial image fusion using convex optimization, in: M. Goesele, S. Roth, A. Kuijper, B. Schiele, K. Schindler (Eds.), DAGM-Symposium, in: Lecture Notes in Comput. Sci., vol. 6376, Springer, 2010, pp. 303–312.
[31] D. Labate, W.-Q. Lim, G. Kutyniok, G. Weiss, Sparse multidimensional representation using shearlets, in: Wavelets XI, San Diego, CA, 2005, in: SPIE Proc., vol. 5914, SPIE, Bellingham, WA, 2005, pp. 254–262.
[32] A. Llagostera Casanovas, G. Monaci, P. Vandergheynst, R. Gribonval, Blind audiovisual source separation based on sparse representations, IEEE Trans. Multimedia 12 (5) (August 2010) 358–371.
[33] Y.M. Lu, M.N. Do, A theory for sampling signals from a union of subspaces, IEEE Trans. Signal Process. 56 (6) (June 2008) 2334–2345.
[34] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online learning for matrix factorization and sparse coding, J. Mach. Learn. Res. 11 (March 2010) 19–60.
[35] S. Mallat, Zero-crossings of a wavelet transform, IEEE Trans. Inform. Theory 37 (4) (July 1991) 1019–1033.
[36] S. Mallat, A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way, 3rd edition, Academic Press, 2008.
[37] S. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries, IEEE Trans. Signal Process. 41 (1993) 3397–3415.
[38] S. Nam, M. Davies, M. Elad, R. Gribonval, Cosparse analysis modeling – uniqueness and algorithms, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2011, Prague, Czech Republic, May 2011.
[39] S. Nam, M. Davies, M. Elad, R. Gribonval, Recovery of cosparse signals with greedy analysis pursuit in the presence of noise, in: The Fourth International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, San Juan, Puerto Rico, December 2011.
[40] B. Natarajan, Sparse approximate solutions to linear systems, SIAM J. Comput. 25 (2) (1995) 227–234.
[41] D. Needell, J.A. Tropp, CoSaMP: iterative signal recovery from incomplete and inaccurate samples, Commun. ACM 53 (12) (2010) 93–100.
[42] M.D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, M.E. Davies, Sparse representations in audio and music: from coding to source separation, Proc. IEEE 98 (6) (June 2010) 995–1005.
[43] J. Portilla, Image restoration through l0 analysis-based sparse optimization in tight frames, in: Proceedings of the 16th IEEE International Conference on Image Processing, 2009, pp. 3865–3868.
[44] H. Raguet, J. Fadili, G. Peyré, Generalized forward–backward splitting, Technical report, preprint Hal-00613637, 2011.

[45] L. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, Phys. D 60 (1992) 259–268.
[46] I.W. Selesnick, M.A.T. Figueiredo, Signal restoration with overcomplete wavelet transforms: comparison of analysis and synthesis priors, in: Proceedings of SPIE, vol. 7446, Wavelets XIII, August 2009.
[47] K. Skretting, K. Engan, Recursive least squares dictionary learning algorithm, IEEE Trans. Signal Process. 58 (4) (2010) 2121–2130.
[48] J.-L. Starck, E.J. Candès, D.L. Donoho, The curvelet transform for image denoising, IEEE Trans. Image Process. 11 (11) (November 2002) 670–684.
[49] J.-L. Starck, M. Elad, D.L. Donoho, Redundant multiscale transforms and their application for morphological component analysis, Adv. Imaging Electron Phys. 132 (2004).
[50] J.-L. Starck, F. Murtagh, E.J. Candes, D.L. Donoho, Gray and color image contrast enhancement by the curvelet transform, IEEE Trans. Image Process. 12 (6) (2003) 706–717.
[51] J.-L. Starck, F. Murtagh, M.-J. Fadili, Sparse Image and Signal Processing – Wavelets, Curvelets, Morphological Diversity, Cambridge University Press, 2010.
[52] J.F. Sturm, Using sedumi 1.02, a MATLAB toolbox for optimization over symmetric cones, Optim. Methods Softw. 11–12 (1999) 625–653.
[53] V.N. Temlyakov, Weak greedy algorithms, Adv. Comput. Math. 12 (2–3) (2000) 213–227.
[54] J.A. Tropp, Greed is good: algorithmic results for sparse approximation, IEEE Trans. Inform. Theory 50 (2004) 2231–2242.
[55] J.A. Tropp, Just relax: convex programming methods for subset selection and sparse approximation, IEEE Trans. Inform. Theory 51 (2006) 1030–1051.